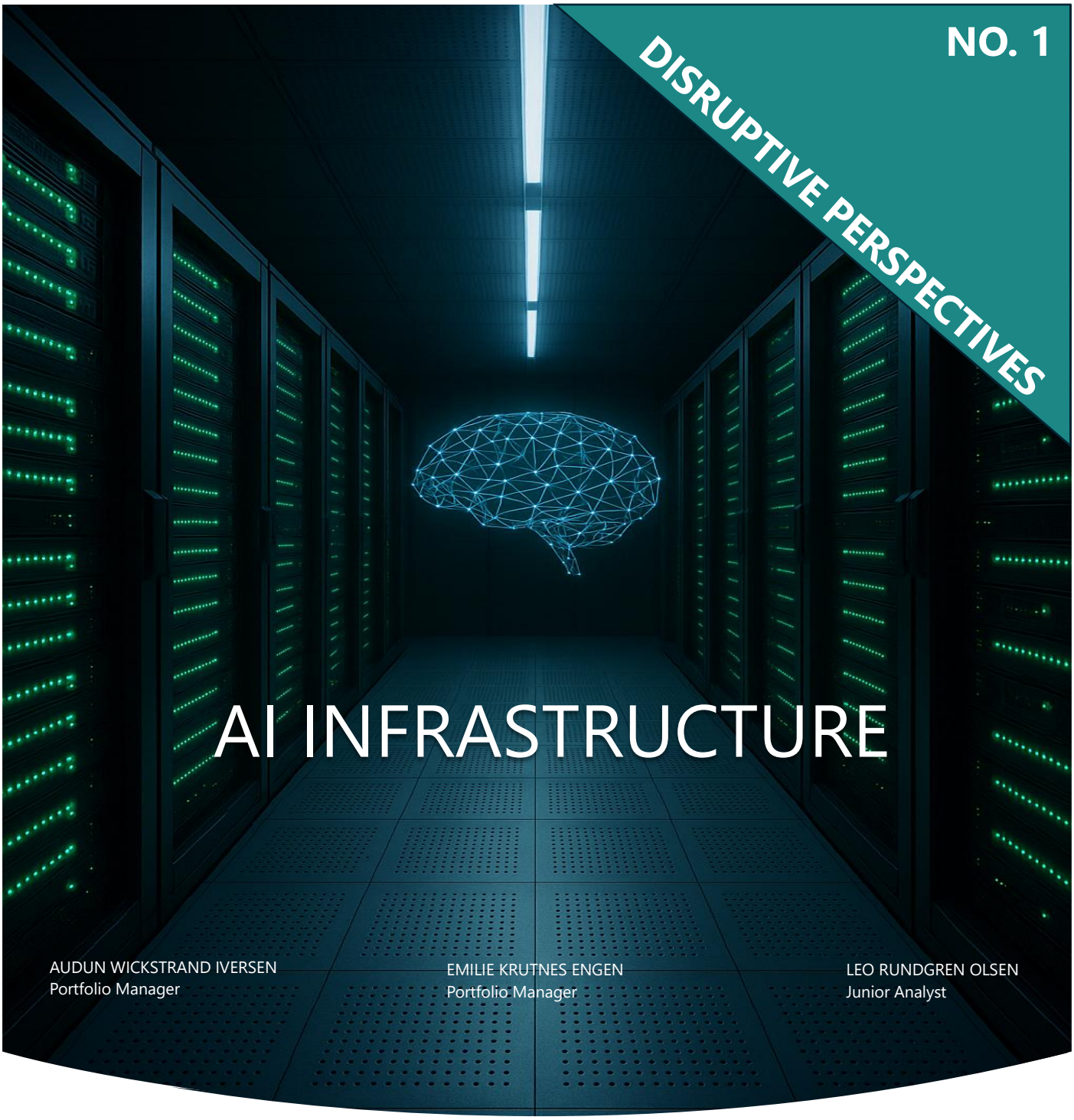


NO. 1

DISRUPTIVE PERSPECTIVES

AI INFRASTRUCTURE

AUDUN WICKSTRAND IVERSEN
Portfolio Manager

EMILIE KRUTNES ENGEN
Portfolio Manager

LEO RUNDGREN OLSEN
Junior Analyst

The purpose of Disruptive Perspectives:

When we analyze different themes, we spend a lot of time using many tools: quarterly reports, analyses, company dialogue, company visits, Excel, calculators and language models. We often make small notes, and sometimes larger notes that we think of as perspectives. We are old enough to know that there are rarely absolute truths, often only different perspectives.

Disruptive Perspectives has only one purpose: to share our perspectives on themes that shape our future. These are not academic papers, encyclopedia entries or recommendations to do, buy or sell anything. Just good old-fashioned information sharing to show how we view different themes at the time of publication. Perspectives do not become smaller when shared, perhaps rather larger. With that as our starting point, enjoy the journey through our perspectives.

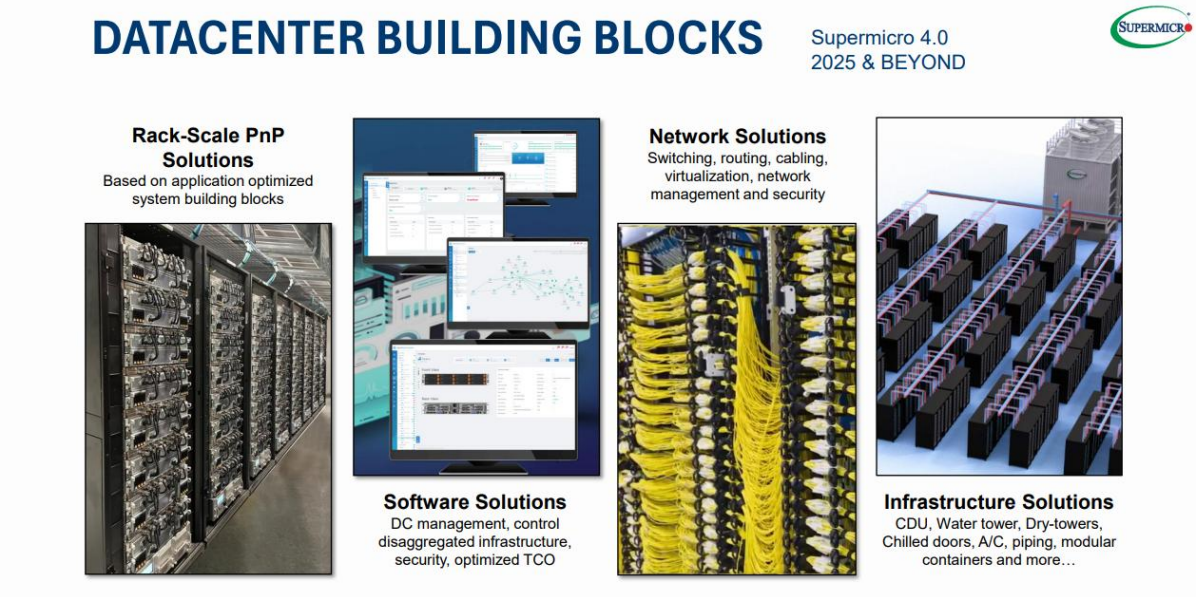
Contents

Disclaimer 1
What is AI infrastructure? 2
The vision 3
Enablers 4
Categorization 7
AI factories 10
What is an AI factory? 10
CUDA and Dojo: 14
Tokens - The new currency of AI 18
Federated Learning 21
What is Edge AI? 24
What are neuromorphic chips? 27
What are Spiking Neural Networks? 31
Tesla and NVIDIA 43
Training: interaction and independence 43
Action (Inference): edge in motion 44
NVIDIA's role in training and action in artificial intelligence 47
SNNs and the rise of humanoids 50
From cloud to edge: 52
When intelligence moves into products 53
A new feedback loop 54
Where are the largest AI opportunities in 2025? 54

Disclaimer

What is AI infrastructure?

When we talk about artificial intelligence, it is easy to think of apps such as ChatGPT, self-driving cars or medical image diagnostics. But we rarely talk about what makes all of this possible. The foundation on which artificial intelligence (AI) is built. This foundation is called AI infrastructure, and it consists of everything from hardware and software to data centers, networks and the people who develop and operate the technology. Without this infrastructure, we get no AI. Or AGI. Or ASI.



Source: Supermicro Computer Q1 report

What makes AI infrastructure unique is that it is extremely demanding. Where traditional IT infrastructure handles websites or databases, AI infrastructure requires massive computing power, enormous data volumes and continuous development. Modern language models are trained on several hundred billion words and run on thousands of GPUs (Graphics Processing Units) for weeks.

Where data infrastructure could previously be built once and used for years, AI infrastructure must be dynamic. Models become outdated quickly, new datasets are added and algorithms must be updated. Modern AI infrastructure is therefore often cloud-based and modular, with hardware such as specially designed GPUs and TPUs (Tensor Processing Units), and software that enables continuous development through MLOps (machine learning operations).

Despite this, AI infrastructure resembles other infrastructure in some areas: it must be scalable, secure and dependent on fast networks. The difference lies in the intensity and complexity. AI needs data. It is essential for further development and for achieving AGI (Artificial General Intelligence) and ASI (Artificial Super Intelligence).

But AI infrastructure is not only about technology. It is also about how all these components fit together in an ecosystem. This ecosystem includes technical processes such as model training and deployment, physical hardware such as supercomputers and edge chips, software platforms and concrete AI products. All of this works together in a system where small changes in one component can have major consequences for the whole. In other words, AI infrastructure is both technological and systemic, and perhaps the most important invisible engine behind technological progress over the next few decades.

The vision

When you build a motorway, it is not only about asphalt and concrete. It is about what it enables. In this case, movement, trade and contact between people. In the same way, AI infrastructure is not only about servers, data centers and algorithms. It is about creating a platform for future innovation.

The great goal of AI infrastructure is to make advanced artificial intelligence available to everyone. Not only to the technology giants. This democratization makes it possible for startups, researchers, the public sector and individuals to build solutions that were previously unthinkable. Think of climate models, personalized medicine or intelligent learning tools. All of this depends on the underlying infrastructure being available, robust and sustainable.

But to reach this goal, we must solve some fundamental challenges.

- Scale: the infrastructure must handle both enormous datasets and extremely complex models.
- Environment: today's AI training requires enormous amounts of energy, and without better energy efficiency, growth will become unsustainable.
- Accessibility: AI must be usable by smaller actors, not only those with billion-dollar budgets.

- Ethics, security and reliability: we must be able to trust that the systems we build do not discriminate, misinterpret or leak sensitive information.

When we assess how good AI infrastructure is, we therefore need to look broadly: how quickly are the models trained? How much energy do we use per prediction? How accessible is the technology for new actors? How safe and fair are the systems in practice? The success of AI infrastructure cannot be measured only in technical performance. It must also work for society.

Enablers

When we lift our gaze, we see that AI infrastructure stands on the shoulders of six powerful enablers:

First, we have specialized hardware (GPUs) and neuromorphic chips that give systems enough power to learn. Moore's law still plays a role here, but it is now supplemented by new principles such as Amdahl's law and innovations in quantum computing.

Then comes data. Without data, no AI. But a lot of data is not enough. It must be relevant, clean and continuously updated. Edge devices and synthetic datasets are becoming increasingly important to meet this need.

On the software side, we find frameworks such as PyTorch and TensorFlow, and MLOps platforms that allow models to be trained, deployed, monitored and improved in real time.

Cloud platforms give us elastic capacity and global availability, while edge computing gives us low latency and local control. Two approaches that complement each other.

But the technology itself is not enough. Without people who develop, maintain and challenge it, we will not get far. AI requires interdisciplinary expertise, from coding and statistics to ethics and domain insight. Here, we need to think like a learning organization, where knowledge is continuously updated and shared.

Finally, economic and regulatory conditions set the framework. Public support, tax incentives and risk capital are crucial, but so are regulations such as GDPR, which ensure that AI is developed and used responsibly.

What might AI-generated regulations look like?

A typical AI-generated regulation would probably be highly detailed and technical. Using Natural Language Processing (NLP) and machine learning models, AI would propose regulations based on existing rules, technical specifications and public input. An example could be:

"AI systems classified as high risk (defined as an error rate above 0.1% in critical applications) shall log all decisions in an auditable format and undergo annual bias testing."

Such proposals would be highly precise, but could easily lack the necessary flexibility, cultural understanding and local adaptation. These are elements that require human judgment.

The optimal solution: humans supported by AI

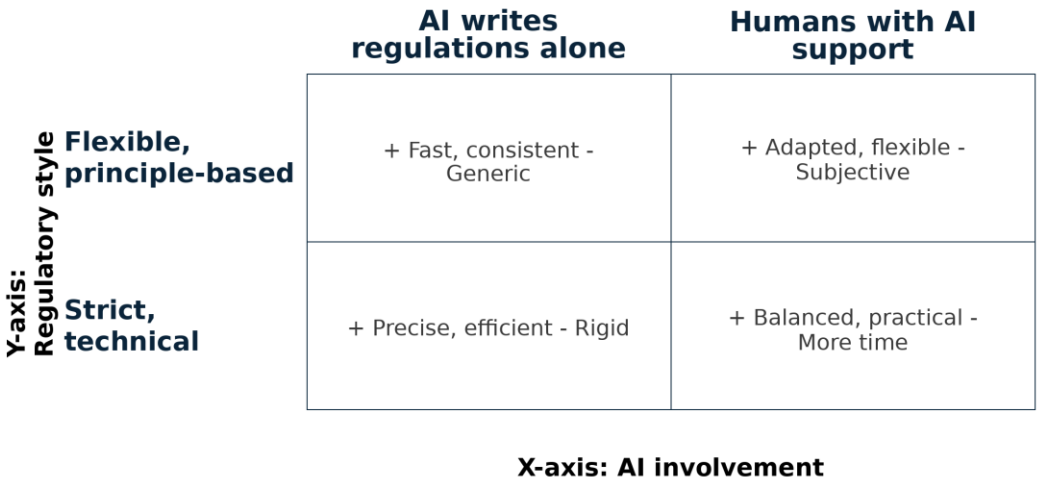
Even though AI should not be given full responsibility for regulatory work, that does not mean we should ignore its potential. The best approach is to use a hybrid model, where AI functions as a powerful support tool.

A combination of AI and human judgment can produce precise and data-driven drafts from AI, which are then refined by people. Humans can ensure that regulations take ethical, cultural and societal dimensions into account. This secures a balance between technical precision and practical applicability, while also preserving democratic legitimacy.

We can make this clearer with a 2x2 matrix showing what role AI should play depending on the nature of the regulation:

Postcards From The Future

Regulations



These six factors do not operate alone; they are connected in a dynamic system. When one of them fails, the effect propagates through the entire ecosystem. This is systems theory in practice: everything is connected.

We must therefore build an infrastructure that is powerful and flexible, ethical and efficient, human-driven and machine-augmented. And we must do it in a way that is open to innovation, while also being aware of the consequences.

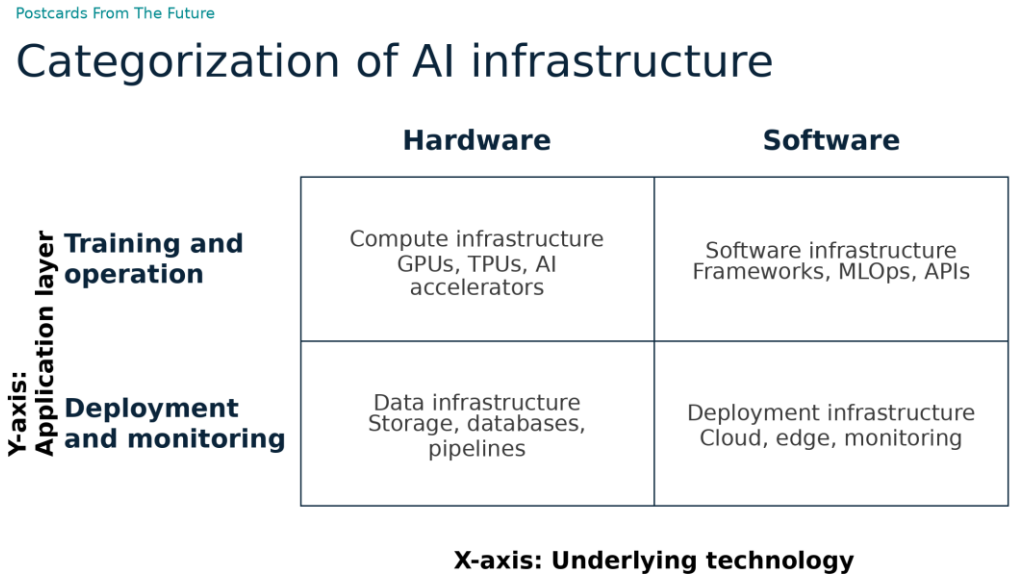
With this as our starting point, we move on to the different types of AI infrastructure, and how they are organized technologically and practically in the world of today and tomorrow...

Categorization

As we now move deeper, we dive into different types of AI infrastructure and the key technologies that form the backbone of modern AI. The goal is to provide an understandable and coherent explanation of how these technologies fit together.

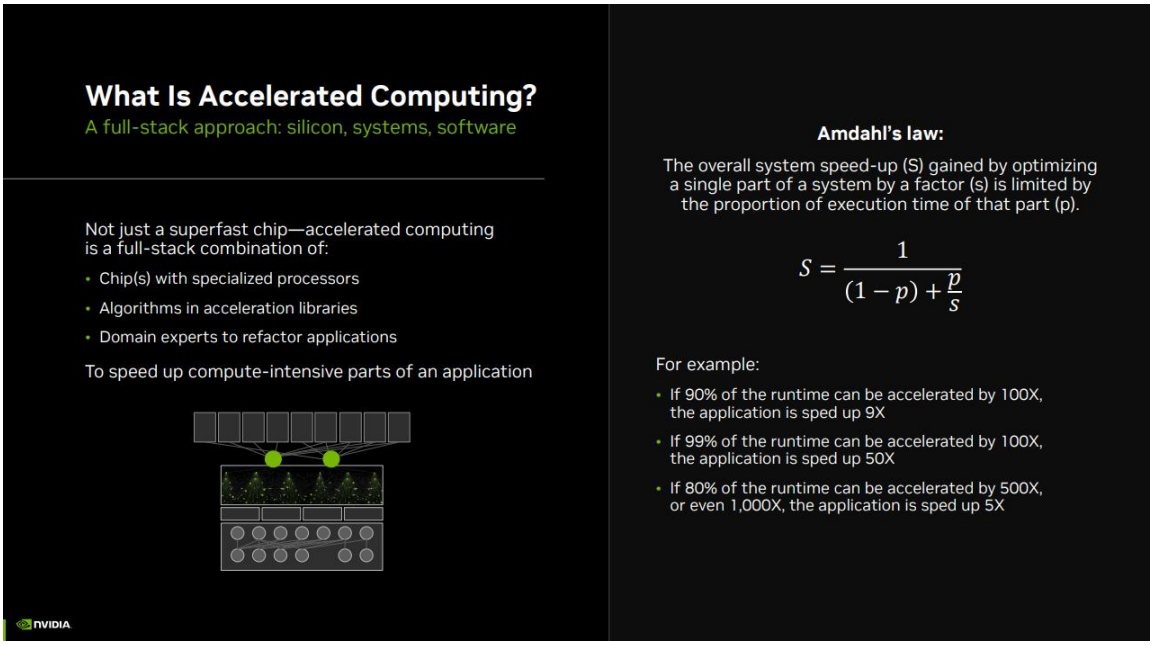
AI infrastructure consists of hardware, software, data storage and networks that together enable the development, training, operation and maintenance of advanced AI models. This integrated system naturally divides into four main areas:

- *Compute infrastructure*
- *Data infrastructure*
- *Software infrastructure*
- *Deployment infrastructure*



Compute infrastructure = the engine

This is the heart of AI: specialized processors such as GPUs (e.g. NVIDIA H100) are built for massive number crunching. In addition, there are neuromorphic chips such as Intel Loihi, which mimic the brain's efficiency, and edge hardware such as NVIDIA Jetson and ARM devices that make AI runnable directly on devices such as robots or sensors. Larger systems, such as NVIDIA DGX clusters, combine these to train powerful models. This follows the trends we know from Moore's and Amdahl's laws: ever more power and a growing need for parallel performance.



Source: NVIDIA Investor Relations

Data infrastructure is the fuel behind intelligence

Without good data, computing power is of little use. Here we are talking about data pools and object storage such as Amazon S3, distributed databases and platforms such as Snowflake. Data is often cleaned and structured through Spark, Pandas and dedicated tools, and can come from everything from IoT sensors and public open datasets to synthetically generated data. Good data governance, with anonymization, quality assurance and GDPR compliance, helps avoid the GIGO, "garbage in, garbage out", principle. In short: without clean fuel, the AI factory stops.

Software infrastructure makes it possible to build complex systems by "hiding" the technical layer underneath, so developers can focus on what they are building, not how everything works in detail. It uses principles from abstraction, meaning the simplification of something complex so more people can contribute.

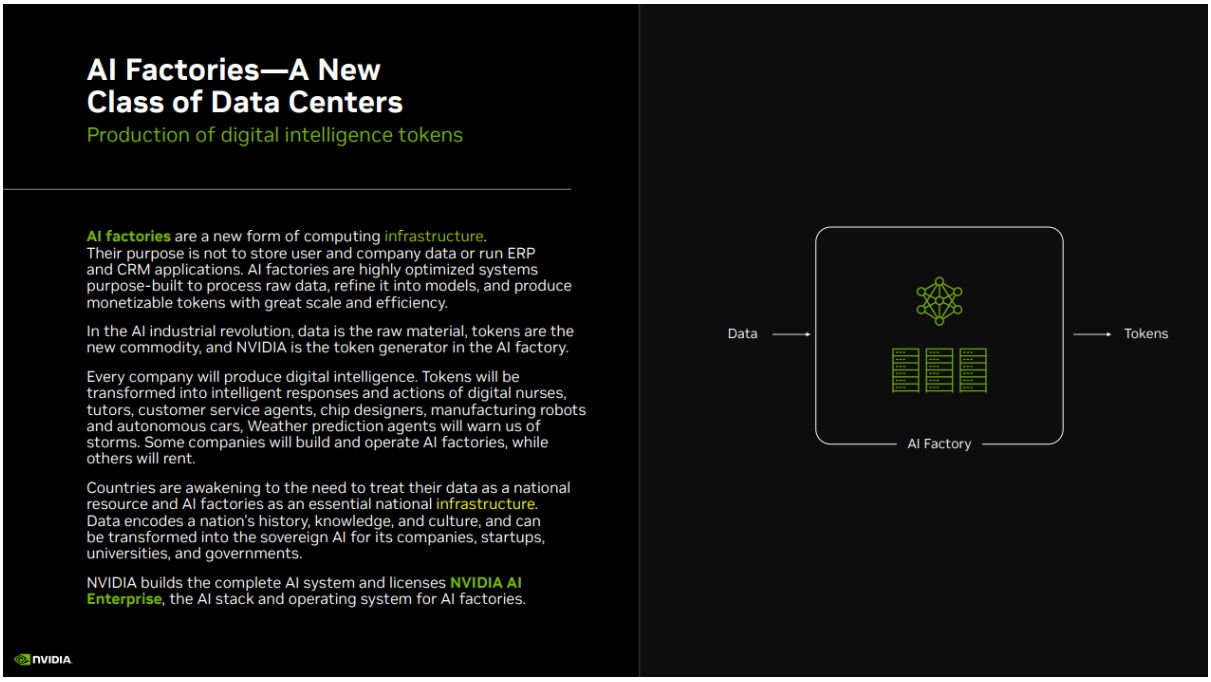
At the same time, it facilitates innovation by being accessible through the use of open-source code. This means that anyone can see, use and improve the software, creating a shared development platform where ideas and solutions grow faster. Frameworks such as TensorFlow, PyTorch and JAX make complex mathematics accessible to developers. Tools such as MLflow and Kubeflow handle the full lifecycle of models, from training to production. And thanks to cloud services such as AWS SageMaker, Azure Machine Learning and Google Vertex AI, small actors can access advanced software without a large upfront cost. Libraries such as NumPy and Dask make data processing smooth and fast. The strength lies in the fact that much of this is open source.

Deployment infrastructure builds on principles from DevOps (continuous integration and deployment), but is adapted to AI through what is called MLOps (machine learning operations). In short: DevOps (Development Operations) makes it easy to update apps regularly. MLOps makes it possible to update AI models regularly and ensure they still work as intended, learn from new data, and can be managed and monitored in operation. This means that AI models are not trained once and then finished. They are maintained, improved and redeployed continuously, just like software programs. Once a model has been trained, it must work in practice. This is where cloud platforms come in, adapting to the load. For mobile applications, technologies such as TensorFlow Lite and ONNX Runtime are used so AI can run directly on devices. APIs and SDKs such as the OpenAI API or xAI provide simple integration, while Prometheus and Grafana provide monitoring and continuous maintenance through MLOps practices.

With this knowledge in mind, we can move on to understand what AI factories, tokens, CUDA, training, inference and edge AI involve, and how these elements affect the development of AI infrastructure in practice.

AI factories

These factories are at the core of modern AI infrastructure. They are not factories in the traditional sense, but data centers built for one thing: to produce intelligence. The term was first popularized by NVIDIA and describes facilities that combine enormous amounts of compute with advanced software to train and operate AI models at scale. In practice, these factories convert data and energy into models such as language models, self-driving vehicle systems or industrial robots.

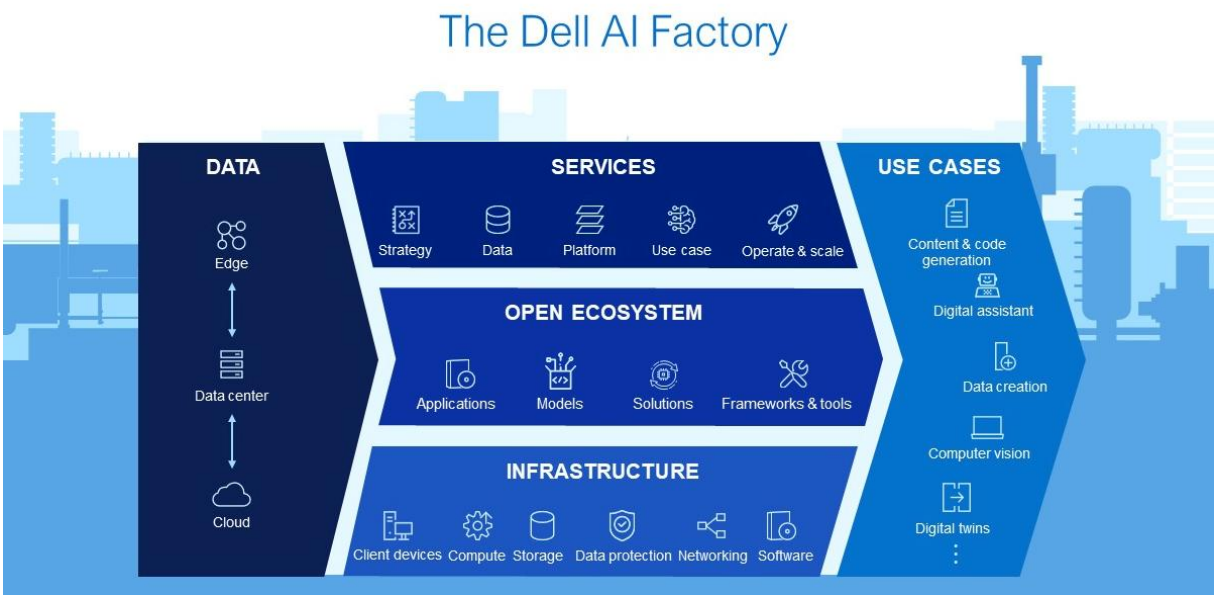


Source: NVIDIA IR

What is an AI factory?

An AI factory is a term that describes highly specialized, scalable computing facilities designed to develop, train and operate artificial intelligence on an industrial scale. According to NVIDIA, AI factories are massive GPU clusters that function as “factories” for generating AI models, applications and intelligent solutions, where data and energy are the raw materials and AI tokens, such as predictions or decisions, are the product. These facilities are often cloud-based or hybrid data centers that combine powerful hardware, advanced software and

data governance to support AI workflows such as deep learning, natural language processing and autonomy.



Source: https://www.dell.com/zh-tw/blog/digital-assistant-solutions-update-with-dell-and-nvidia/?utm_source=chatgpt.com

AI factories are crucial for enabling technologies such as self-driving cars, humanoids and AI agents, including those that use federated learning in cars, which we will look at later.

AI factories are built from many different technologies that work together:

- **Hardware: It starts with powerful chips such as NVIDIA H100 and Google TPUs, and extends to specialized chips such as Tesla's Dojo D1 and neuromorphic chips such as BrainChip Akida. These handle enormous amounts of data and compute operations.**
- **Software: Frameworks such as TensorFlow and PyTorch are used for model development, while MLOps tools such as MLflow and Kubeflow help automate the processes. Tools such as NVIDIA TensorRT are used to optimize performance.**
- **Edge: When AI runs locally, as in cars or humanoid robots, hardware such as NVIDIA Jetson and software such as ONNX Runtime are used.**
- **Federated learning: Models can be trained locally on devices and updated centrally, a technique particularly associated with Tesla.**

- **Synthetic data: When real data is not available, realistic datasets are simulated using platforms such as NVIDIA Omniverse and CARLA.**

Data is the most important raw material in an AI factory. Raw data must be collected, "cleaned", labeled and analyzed. Tools such as Apache Spark, Pandas and Databricks are used for this. In many cases, data is analyzed in real time, as in cars or smart cities, where streams of data are continuously sent in. Synthetic data created in simulators is also used to supplement or replace real-world datasets.

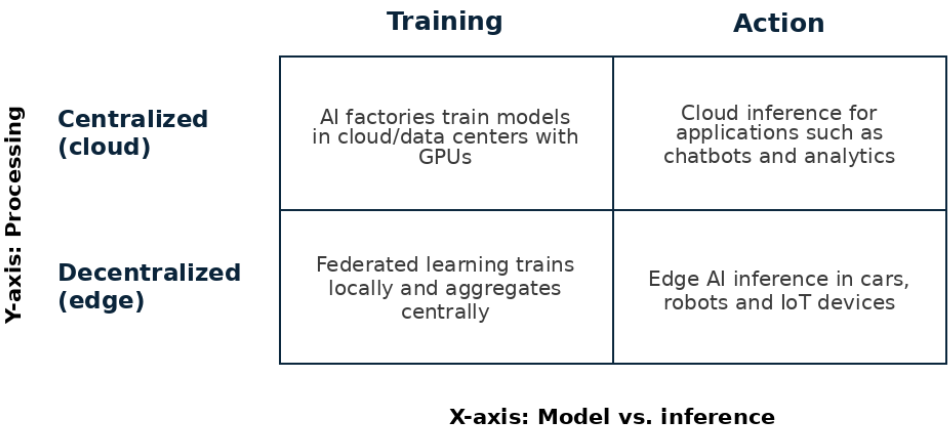
The factories are part of a larger network of actors and industries that collaborate:

- **Hardware suppliers such as NVIDIA, Intel and Tesla build the chips.**
- **Cloud platforms such as AWS, Azure and Google Cloud provide training infrastructure and API services.**
- **Software companies such as Hugging Face and Databricks build models, frameworks and analytics tools.**
- **The automotive industry, led by Tesla, Waymo, WeRide, Pony.ai and others, uses AI factories to train autonomous driving systems.**
- **Robotics companies such as Boston Dynamics, Agility, 1X, Unitree, Apptronik, Tesla and Figure AI train humanoids in the factories.**
- **Energy companies such as Schneider, Emerson, EOSE, ADSE and IREN contribute green power to make operations sustainable.**
- **Academia and research institutions such as MIT, Stanford and DeepMind drive the development of the next generation of AI.**

We can divide AI factories according to where they operate (cloud vs. edge) and what they do (training vs. inference):

Postcards From The Future

The AI factory ecosystem



Explanation of the quadrants:

1. Training + centralized:

   AI factories train models in the cloud or in data centers using GPUs or specialized chips. Example: NVIDIA DGX for GPT-4, Tesla Dojo for FSD.

   Actors: NVIDIA, Tesla, Google, AWS.

2. Action + centralized:

   Cloud-based inference for applications such as chatbots or analytics. Example: Azure AI for inference, NVIDIA Triton.

   Actors: Microsoft, Google, NVIDIA.

3. Training + decentralized:

   Federated learning trains models on edge devices and aggregates them in AI factories. Example: Tesla cars train FSD locally.

   Actors: Tesla, Waymo, Google.

4. Action + decentralized:

   Description: Edge AI inference in cars, humanoids or IoT devices. Example: NVIDIA DRIVE Orin, BrainChip Akida.

   Actors: NVIDIA, Tesla, Intel, BrainChip.

AI factories are ecosystems. Every component, from hardware and software to data security and energy source, plays a role. Systems theory helps us understand how these parts fit

together. Innovation theory explains why actors such as NVIDIA and Tesla succeed in disrupting entire industries.

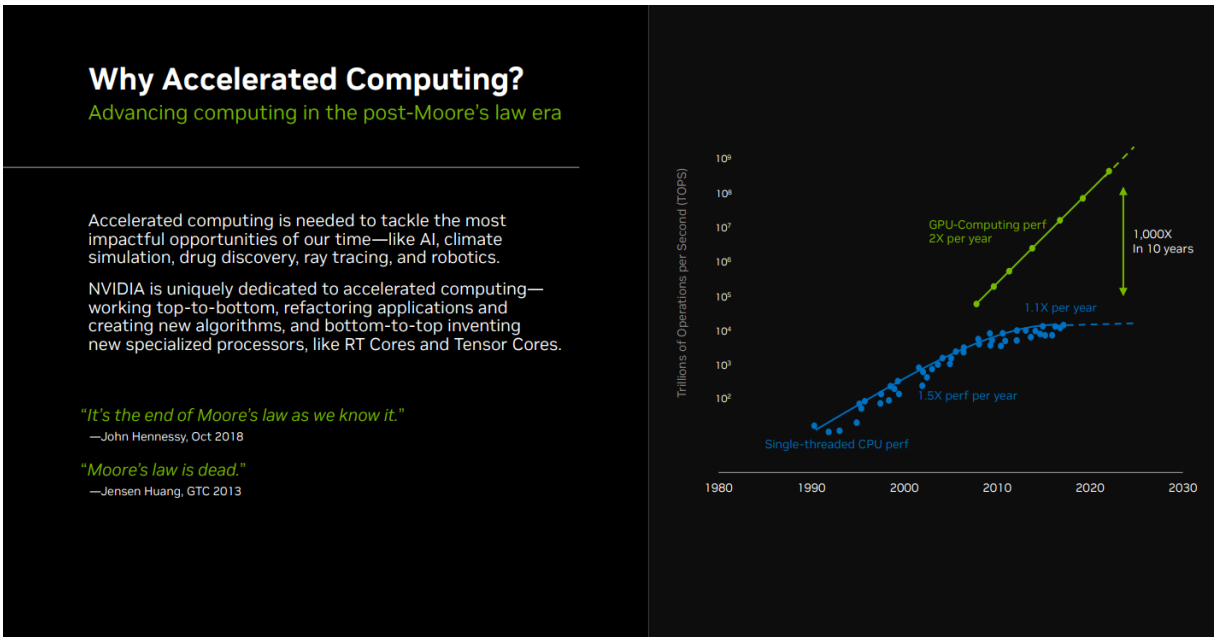
CUDA and Dojo:

Two of the most central technologies in today's AI infrastructure are CUDA and Dojo. They represent two different strategies for solving the same challenge: how to train and use advanced AI models efficiently. While CUDA is the industry standard developed by NVIDIA and used broadly across industries, Dojo is Tesla's internally developed solution, specialized for its video-based AI models used in self-driving cars and humanoid robots. Together, they illustrate both collaboration and competition in the development of AI factories.

What is CUDA?

CUDA (Compute Unified Device Architecture) is NVIDIA's platform for general-purpose computing on GPUs. In practice, this means that developers can use NVIDIA's graphics processors to perform demanding calculations that previously had to be done by CPUs. This makes CUDA ideal for AI training, where large amounts of data and complex calculations must be handled in parallel.

CUDA is deeply integrated into the AI ecosystem through support for popular frameworks such as TensorFlow and PyTorch (SNNs), and offers optimized libraries such as cuDNN for deep learning and cuBLAS for linear algebra. In powerful systems such as NVIDIA DGX and H100 GPUs, with tens of thousands of cores, CUDA accelerates the training of large language models, neural networks and simulations.

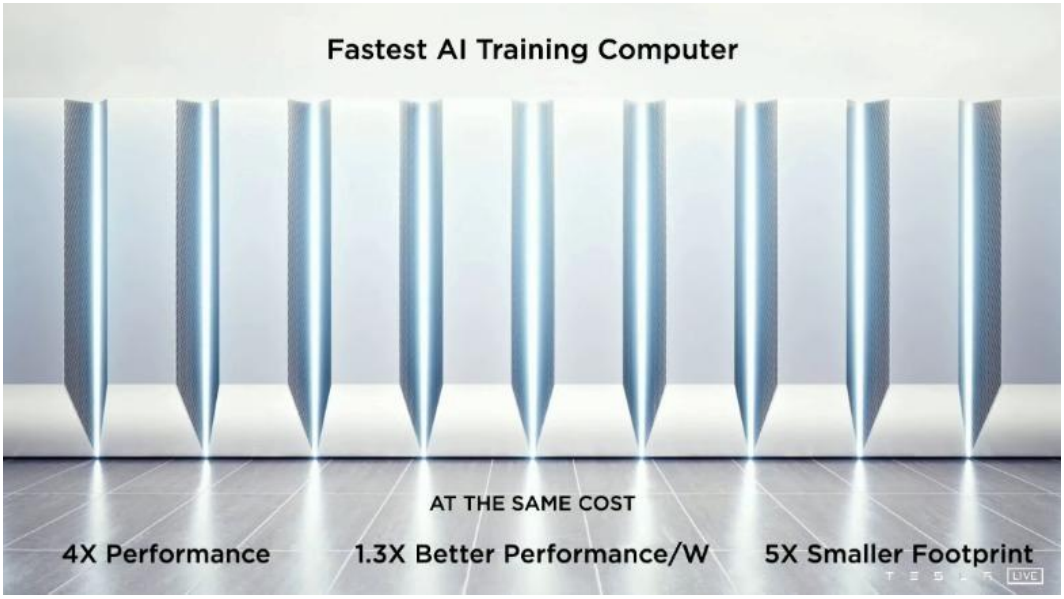


Source: NVIDIA IR

In inference, meaning the actual use of models after training, CUDA enables fast response in both the cloud and edge devices. With tools such as TensorRT, Tesla's self-driving cars can, for example, perform image recognition in under 10 milliseconds. This is crucial in applications where every millisecond counts.

What is Dojo?

This is Tesla's custom-built AI supercomputer, designed for one main task: training its own models for autonomous driving and robotics. Unlike CUDA, which is broadly applicable, Dojo is tailored to Tesla's video-based workflow and vast amounts of vehicle fleet data.



Source: Tesla AI Day 2022

At the core of Dojo is the D1 chip, developed by Tesla and integrated into large wafers for maximum bandwidth and low latency. A single wafer can deliver up to 9 petaFLOPS and handle petabytes of video recordings. These are ideal conditions for training complex models that must interpret traffic, anticipate events and plan driving patterns.

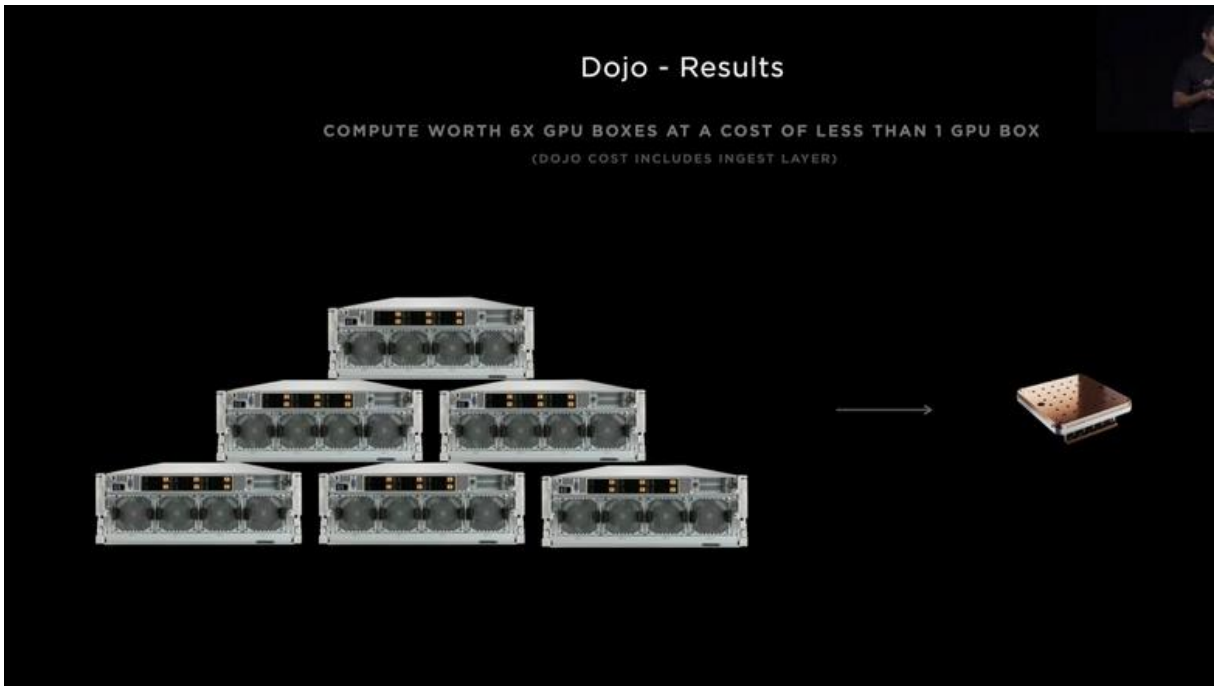
Dojo is tightly integrated into Tesla's federated learning system. Here, insights from millions of cars are collected locally and sent back to Dojo, which updates the models without raw data leaving the cars. In this way, the system improves continuously while preserving privacy.

Interaction

CUDA and Dojo are often used in combination in Tesla's universe. Today, CUDA is still used for much of the training and almost all inference, especially on Tesla's edge devices such as HW3 and HW4. At the same time, Dojo is being developed to take over more of the training work, especially for video-based analysis, and over time to reduce dependence on NVIDIA.

This illustrates an interesting dynamic: while Tesla collaborates with NVIDIA and uses its technology, it is also building a specialized solution that could become a competitor over the longer term.

In decentralized AI workflows such as federated learning, both CUDA and Dojo play important roles. CUDA accelerates local training in the cars, while Dojo functions as a central node that collects and processes the updates. This combination makes it possible to train models in an efficient, energy-efficient and secure way. Ideal for large-scale autonomy.



Source: Tesla AI Day 2022

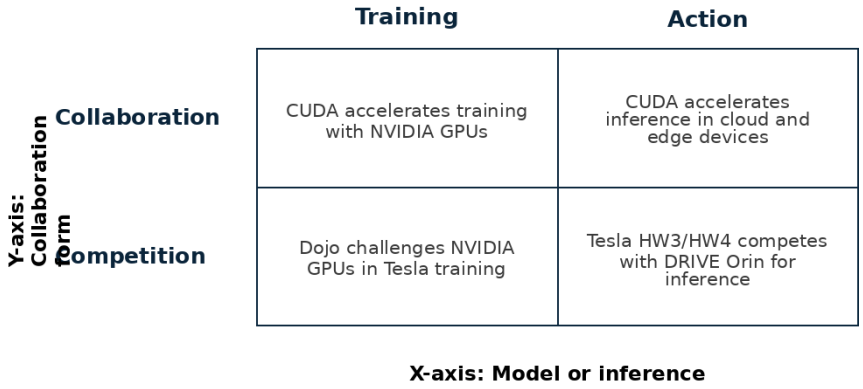
CUDA and Dojo can be seen as different subsystems in the broader AI ecosystem: CUDA as a general solution with broad support, and Dojo as a specialized engine for Tesla's vertically integrated infrastructure. Innovation theory points to how CUDA once revolutionized AI by making GPU computing accessible, while Dojo represents a new wave of specialized, company-internal solutions.

The sustainability perspective is also relevant: Dojo is designed to be more energy-efficient than traditional GPU clusters and supports Tesla's goal of reducing costs and carbon footprint. Privacy theory also plays a role, especially when fleet data is kept locally and only model updates are shared.

We can summarize the differences and roles as follows:

Postcards From The Future

CUDA and Dojo in training and action



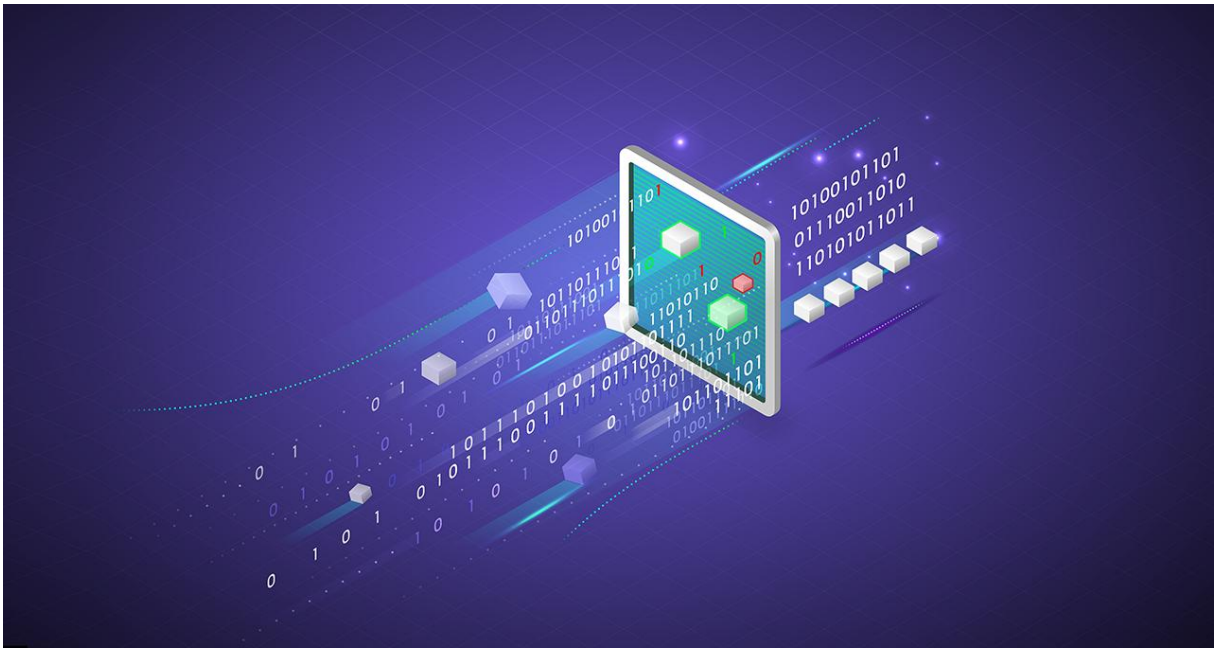
In sum, CUDA and Dojo represent two different approaches to AI infrastructure. One is universal and open to everyone; the other is narrow and targeted. But both are crucial for understanding how AI is actually produced, and how the autonomous systems of the future are developed in practice.

Tokens - The new currency of AI

Jensen Huang, the man in the leather jacket at NVIDIA, positions the company as a cornerstone of a new industrial revolution. AI factories are data-driven systems built on NVIDIA's GPUs and AI platforms such as Blackwell. These factories produce tokens, a new unit of economic value. Electricity enabled industrial production, and the internet drove the information economy. Huang sees AI infrastructure as the next wave, where intelligence, measured as tokens, becomes a universal resource that transforms economies. We start calmly and end up in a 2x2 matrix.

Huang says NVIDIA is no longer just a chip company, but a provider of infrastructure for AI-driven economies. This includes GPUs, AI platforms and software such as CUDA and Blackwell that power everything from self-driving cars to humanoid robots. NVIDIA has traditionally sold GPUs, which have been used especially in gaming and powerful PCs, and so far have become the building blocks of the new data centers.

The AI factories we discussed process enormous amounts of information to train and run AI models. Instead of producing clothes or bread, these “factories” produce tokens. Tokens are a new unit of AI output, such as text, images, decisions or actions generated by models. One token can be a word in a text, a pixel in an image or a decision. Models such as ChatGPT generate text tokens, while the software in Tesla’s robotaxis produces decision tokens in real time.

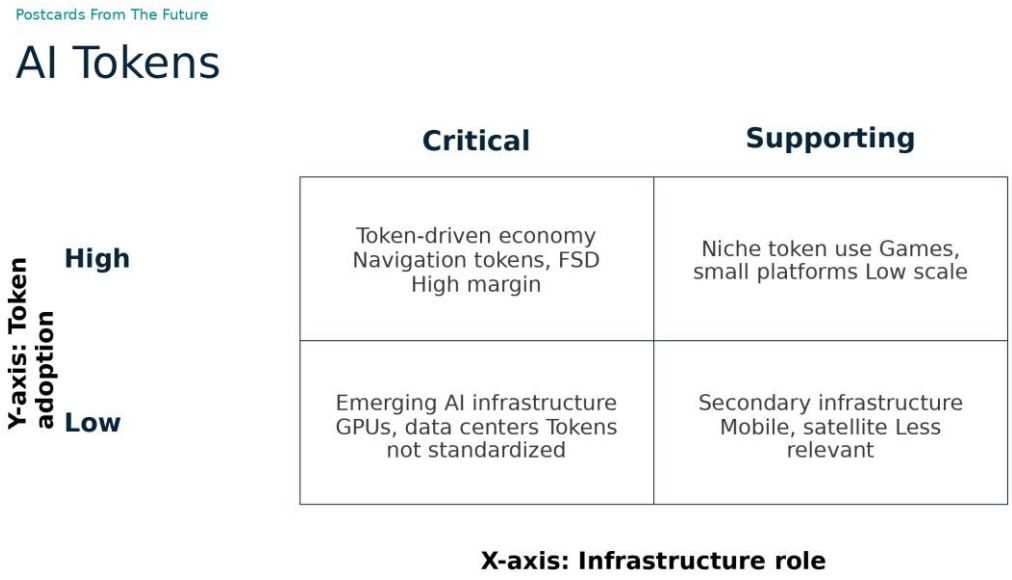


Source: Explaining Tokens — the Language and Currency of AI | NVIDIA Blog[*Explaining Tokens — the Language and Currency of AI | NVIDIA Blog*]

Huang suggests that tokens could become a new currency for productivity, similar to how kilowatt-hours measure electricity or gigabytes measure data traffic. For example, an AI agent that controls a robotaxi could produce “decision tokens” per hour. Huang’s vision of measuring economic output in “tokens per hour” suggests a world where AI productivity is quantified directly, much like we measure factory output in units per hour. For example, a humanoid robot could produce 1,000 decision tokens per hour in a factory, or 500 emotional tokens in elderly care. Establishing a general unit of measurement in AI will probably become harder as we move from language models into other domains. Extending tokens to decision tokens for FSD and action tokens for robots is logical, but requires standardization. For example, a robotaxi could produce “navigation tokens” per kilometer.

If tokens become a unit of measurement, they can be priced directly, for example at $0.01 per token for text generation or $0.10 per decision token for self-driving. Business models around tokens could create high-margin revenues, similar to Apple's App Store.

In our 2x2 matrix, we call the X-axis the role of the infrastructure, distinguishing between critical and supporting, and the Y-axis the degree of token adoption, from high to low:



We call the first quadrant "Emerging AI infrastructure", where critical infrastructure and low token adoption meet. NVIDIA's AI factories are critical for self-driving cars, robots and AI agents for language models, but tokens are not yet a standard unit of measurement.

We call the second quadrant the "token-driven economy", where critical infrastructure meets high token adoption.
If tokens become standard, such as navigation tokens for FSD or emotional tokens for robots, NVIDIA could become one of the dominant companies in a token-based economy. Margins remain high because of software dominance.

We call the third quadrant "secondary infrastructure", where supporting infrastructure and low token adoption meet.

Technologies such as satellites, mobile phones and drones support AI factories, but are not critical. Tokens are irrelevant here.

In the fourth quadrant, we find “niche token use”. Here, the infrastructure is supporting, but token adoption is high.
Small actors can use tokens in niches, but lack scale. Margins are low because of competition. It becomes a bit like a failed cryptocurrency or game.

Whether the man in the leather jacket’s vision of a new industrial revolution, with AI factories producing tokens, becomes a reality, we do not know. But we are watching, and you should too.

Federated Learning

One of the most exciting breakthroughs in AI infrastructure in recent years is the rise of Federated Learning (FL). This is a new way to train AI models. Not in a centralized data center, but directly on decentralized devices such as smartphones, robots and sensors.

Instead of collecting large amounts of raw data in the cloud, the model is sent out to each individual device. There it is trained locally on the user’s data, and only small updates, such as the model’s weights, are sent back. These are then combined on a central server to improve the global model. In this way, the data remains where it belongs: on the device. This provides better privacy, lower bandwidth use and the possibility of real-time improvement on the device.

How it works - step by step:

1. **The model is distributed: a central server sends an initial model to thousands of devices.**
2. **Local training: each device trains the model on its own data.**
3. **Updates are sent back: only model updates, not data, are returned.**

4. **Global aggregation: a central server combines the updates, often using a method called Federated Averaging.**
5. **New round: the improved model is sent out again, and the process repeats.**

Why is this so important? Federated learning has several unique advantages:

- **Privacy: data never leaves the device, in line with regulations such as GDPR.**
- **Efficiency: less need for heavy data storage and cloud access.**
- **Scalability: the possibility of learning across billions of devices.**
- **Real-time learning: models can be improved continuously, based on actual use.**

This makes the technology particularly useful in Edge AI. In other words, when machine learning happens close to the user, for example in a self-driving car or a robot in healthcare.

Examples

- **Google uses FL in the Gboard keyboard so writing patterns improve the model without anything being sent to the cloud.**
- **Apple uses it in Siri and QuickType.**
- **Tesla combines it with Dojo to improve the FSD model based on data from the vehicle fleet.**

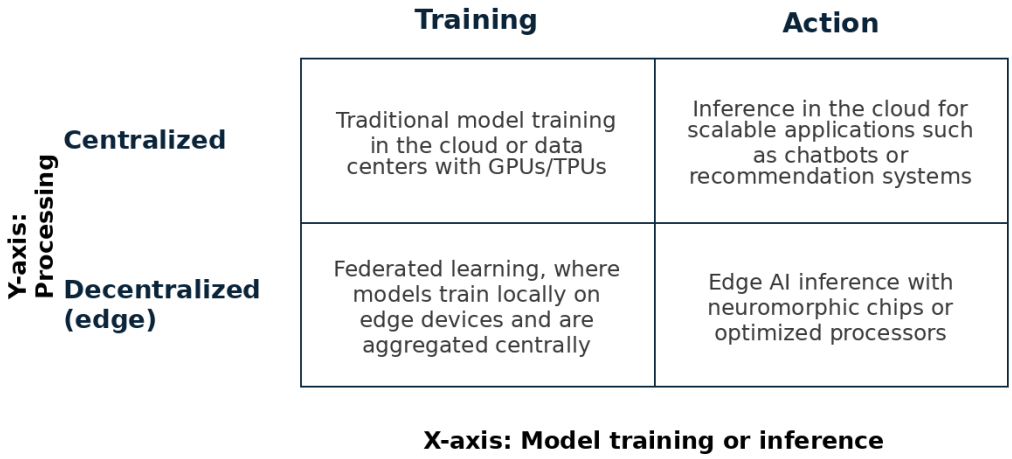
On the hardware side, neuromorphic chips such as Intel Loihi and BrainChip Akida are tailored for such scenarios. They are designed to learn in a power-efficient way and work well on edge devices where compute power and battery are limited.

Traditional AI training takes place centrally in large data centers powered by GPUs and massive datasets. Federated learning turns this on its head. It not only decentralizes learning, but combines it with action. In other words, the device both uses the model (inference) and improves it (training).

We can imagine four types of AI workflows in a 2x2 matrix:

Postcards From The Future

Federated Learning in AI infrastructure



Federated Learning sits in the lower-left corner: decentralized training, close to the user.

Although the technology has significant upside, it also has weaknesses:

- **Variable data quality: heterogeneous data across devices creates challenges for the aggregated model.**
- **Limited hardware: not all devices can train large models efficiently.**
- **Communication: model updates still need to be sent over networks, which is especially challenging in areas without 5G.**
- **Security: even without raw data, it may theoretically be possible to reconstruct sensitive information from model updates. Techniques such as differential privacy are therefore often used.**

Federated learning represents a transition to distributed AI infrastructure, supported by theories from systems thinking, privacy and sustainability. Google introduced the term in 2016, and since then the technology has been driven forward by the need for data security and scalability in a hyperconnected world.

In a future where billions of devices are connected to the network and data is produced everywhere, Federated Learning appears to be a logical path forward for privacy, efficiency and sustainability.

What is Edge AI?

Edge is about bringing intelligence closer to the source. Instead of sending data to large data centers for processing, artificial intelligence runs directly on devices in the field, so-called edge devices. These can be anything from smartphones and drones to sensors in smart cities or robots in a factory. These devices use pretrained AI models to make decisions locally, often in real time, without sending data to the cloud. The result is lower latency, better privacy and lower energy consumption.

Why is this important?

- Real-time decisions: essential for autonomous vehicles and smart cameras that must react within milliseconds.
- Privacy: by keeping data on the device, the risk of data leaks is reduced and requirements from GDPR and similar regulations are met.
- Energy efficiency: avoiding heavy transfers to the cloud reduces power consumption.
- Scalability: makes it possible to handle billions of IoT devices simultaneously without overloading centralized systems.

Edge AI is supported by specialized hardware and software. Neuromorphic chips such as Intel Loihi, BrainChip Akida and IBM TrueNorth mimic the brain and run AI extremely energy-efficiently. At the same time, optimized software tools such as TensorFlow Lite, ONNX Runtime and Lava have been developed to get the most out of low-resource devices.

- Intel Loihi: research-heavy neuromorphic chip for robotics and real-time systems
- BrainChip Akida: commercially oriented and ready for use in IoT devices
- IBM TrueNorth: pioneering in neuromorphic research, but less used commercially

Historical development

Artificial intelligence at the edge is the result of a slow but clear technological shift:

- **1950-1980: centralized mainframes, no Edge AI**
- **1980-2000: neural networks emerge, but the hardware is too weak**
- **2000-2015: the GPU revolution enables powerful cloud-based AI**
- **2015-2020: early frameworks such as TensorFlow Lite make AI on mobile possible**
- **2020-2025: Edge AI grows rapidly, driven by 5G, IoT and energy-efficient chips**

Advantages:

- Very low latency (e.g. <10 ms)
- Minimal energy consumption (milliwatt level)
- Strong privacy
- Handles massive scale

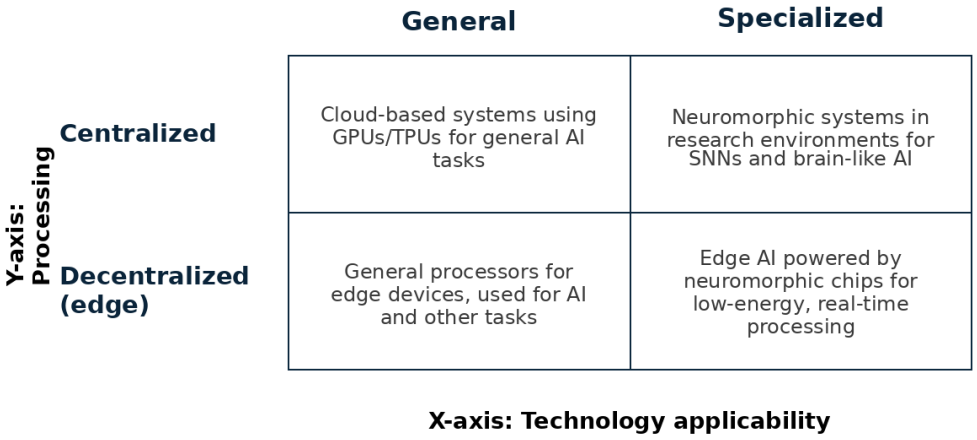
Limitations:

- Limited compute power on the devices
- More difficult to develop and optimize models
- Less mature tools and ecosystem
- Currently limited to specific applications

To place the edge in the bigger picture, we can use a 2x2 matrix:

Postcards From The Future

Edge AI in AI infrastructure



Explanation of the quadrants

1. General and centralized:

   Cloud-based systems that use GPUs/TPUs for general AI tasks. Examples: NVIDIA DGX, Google TPU v4.

   Meaning: dominates large-scale model training, but is energy-intensive.

2. Specialized and centralized:

   Neuromorphic systems in research environments for developing SNNs and brain-like AI. Examples: Intel Hala Point, IBM TrueNorth in research.

   Meaning: important for theoretical progress, but less commercialized.

3. General and decentralized:

   General processors for edge devices, used for both AI and other tasks. Examples: Qualcomm Snapdragon, ARM Cortex.

   Meaning: flexible, but less efficient for AI-specific tasks.

4. Specialized and decentralized:

   Edge AI powered by neuromorphic chips such as Loihi, Akida and TrueNorth for low-energy, real-time processing. Examples: Akida in IoT sensors, Loihi in robotics.

Meaning: drives sustainability and real-time AI, but is limited to niche applications.

This matrix shows that edge today is mainly in the specialized and decentralized category. It targets niche applications, but the potential is enormous, especially for sustainability and real-time systems.

Looking ahead, Edge AI will play an increasingly important role, and technologies such as Akida 2 and Loihi 2 will lift performance. At the same time, 6G and IoT will open the door to new applications in everything from healthcare and industry to robotics and smart cities. This is a change in how we think about intelligence in systems: from central control to local autonomy.

What about the future?

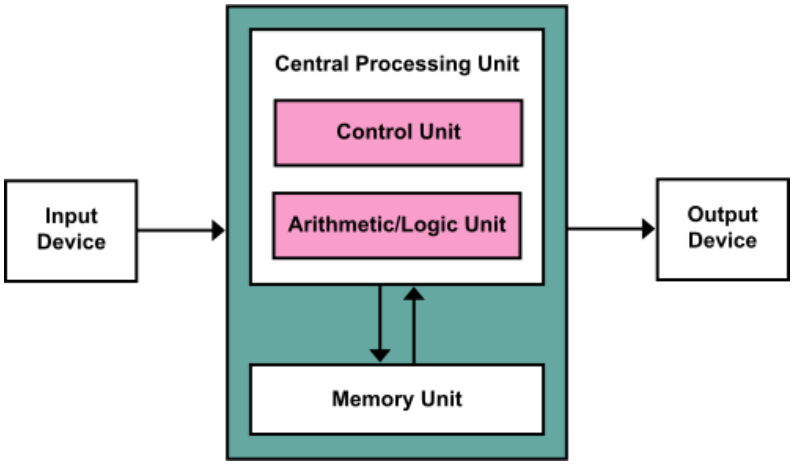
Edge AI will grow with new versions of Akida and Loihi, and with 6G + IoT, healthcare, industry, robotics and smart-city services can take entirely new forms. We are seeing a paradigm shift: from central intelligence to local autonomy, where AI is not only run, but also learns and adapts continuously without exposure. AI infrastructure is a complex, hierarchical system from physical hardware to global distribution and monitoring. It includes powerful AI factories, tokens, decentralized learning and real-time intelligence at the edge.

What are neuromorphic chips?

Neuromorphic chips are a new type of microprocessor inspired by the brain's way of processing information. Instead of using the traditional computer architecture, where processing and memory are separate, neuromorphic chips mimic how neurons and synapses work together in the brain. The result is a distributed, parallel and highly energy-efficient way to perform computation, tailored for tasks such as pattern recognition, sensory processing and low-power AI inference.

The chips use so-called spiking neural networks (SNNs), where "neurons" fire electrical signals ("spikes") only when certain thresholds are reached. Just like in the biological brain. This

event-driven processing makes them far more energy-efficient than traditional processors, which operate continuously with a central clock.

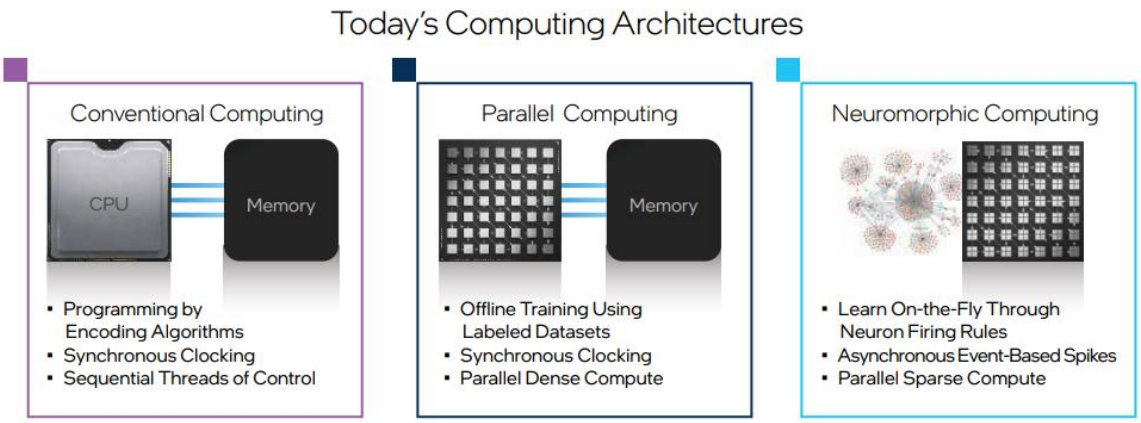


Von Neumann architecture. Source: Von Neumann architecture - Wikipedia[Von Neumann architecture - Wikipedia]

What separates neuromorphic chips from classical computing is that memory and computation happen in the same place, at the neurons. This eliminates the well-known von Neumann bottleneck, where data must be sent back and forth between memory and processor. Combined with local learning and asynchronous processing, this provides very low latency and high efficiency, especially for edge AI and real-time applications.

Examples

- Intel Loihi: research on spiking networks for pattern recognition and robotics, with extremely low power consumption.
- IBM TrueNorth: early prototype with one million neurons, used in image analysis and medical research.
- BrainChip Akida: designed for edge AI, with support for both inference and local learning.



Source: Intel Labs

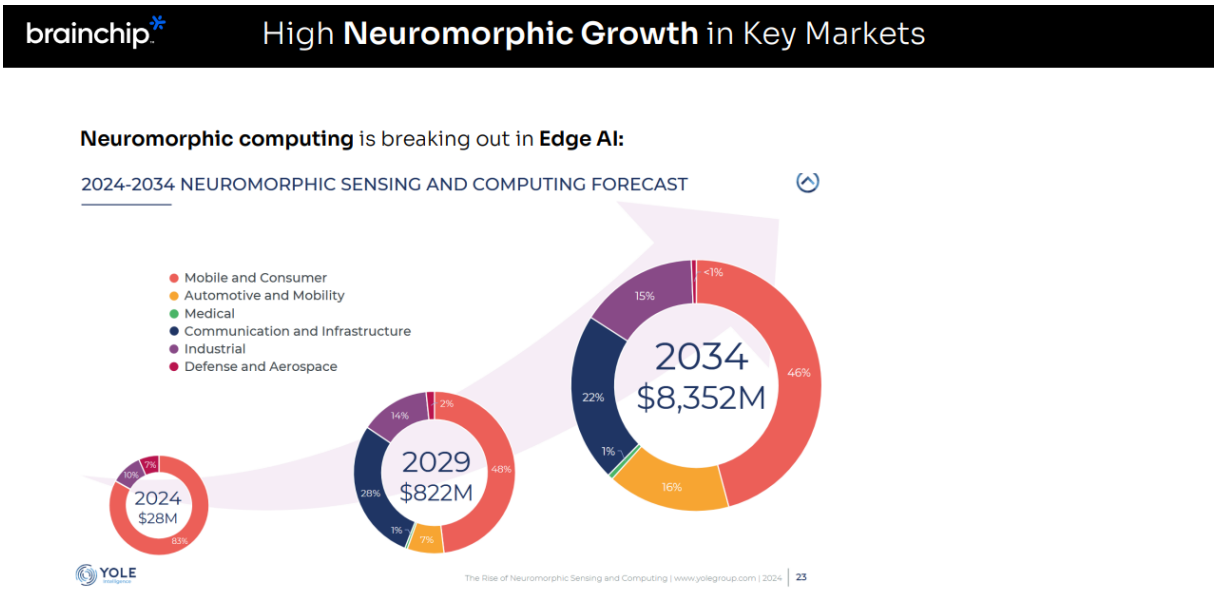
Why is this important?

They have several unique advantages that make them especially well suited for the technology of the future. First, they are extremely energy-efficient. They use power only when something actually happens, a bit like our brain, which activates neurons only when needed. This makes them perfect for small battery-powered devices, such as drones and sensors, that need to be both light and power-efficient.

In addition, they are very good at handling what we call "sparse data" in real time, for example sound, images and movement, where fast response and low latency are important. This means they can react immediately to an audio signal or visual pattern without using much computing power.

Another important advantage is that they can learn locally, directly on the device, without sending data to the cloud. This gives both faster adaptation and better privacy, because the information never leaves the device.

Finally, these chips are tailored for specialized tasks such as pattern recognition or combining information from multiple sensors, tasks that ordinary processors often struggle to perform efficiently.



Source: BrainChip IR

Although neuromorphic chips have great potential, there are also limitations that make them unsuitable for all purposes. First, they are not very flexible. They are not designed for general AI training like traditional GPUs, and therefore fit best with more specialized tasks.

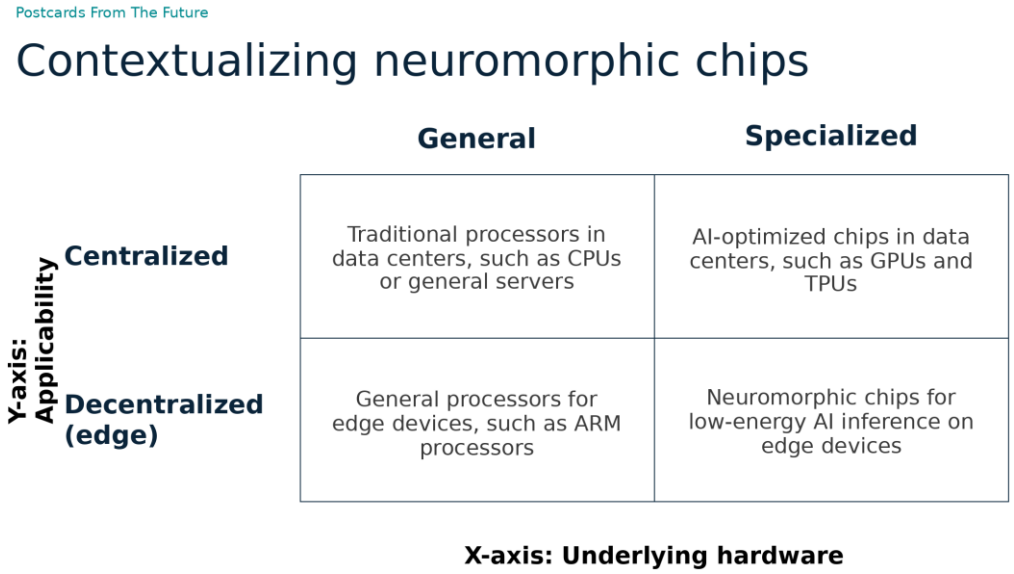
In addition, the ecosystem around the technology is still immature. There are fewer development tools, software libraries and support than in more established AI platforms. This makes development work more demanding.

There is also a high barrier to getting started. Development for neuromorphic chips requires specialized knowledge, both in neuroscience and hardware programming, which limits who can fully exploit the technology.

Finally, the technology is not suitable for large datasets and extensive model training. It works best in edge scenarios, where data is processed locally in small devices, not in large data centers with heavy compute tasks.

Neuromorphic chips belong in a more specialized part of AI infrastructure: closer to the edge of the network than to the data center. They are particularly well suited for applications where real time, power saving and local learning are crucial. In the bigger picture, they contribute to a more decentralized and sustainable AI infrastructure.

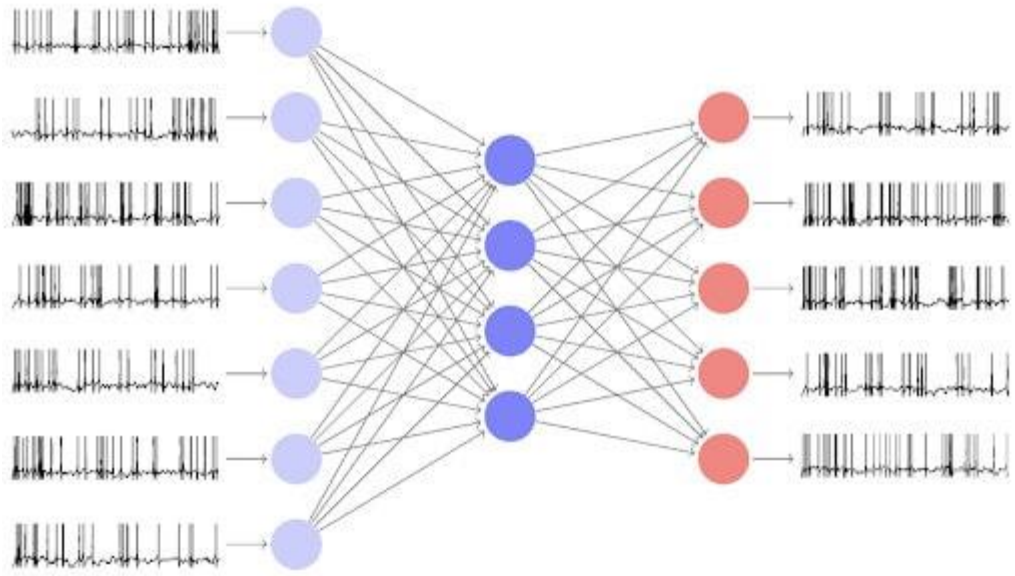
Contextualization in a 2x2 matrix



Neuromorphic chips represent a paradigm shift in how we build AI systems. They integrate computation and memory, reduce energy use and mirror biological principles for learning and signal transmission. From a sustainability perspective, they can play an important role in reducing the carbon footprint of artificial intelligence. At the same time, the technology is still in an early phase and must mature through further research, development and specialized expertise.

What are Spiking Neural Networks?

Spiking Neural Networks (SNNs) are a type of neural network that mimics the brain's way of communicating. Instead of using continuous signals as traditional networks do, SNNs operate with discrete electrical pulses, so-called "spikes", which are sent when an artificial neuron becomes sufficiently activated. This makes them especially suitable for tasks that require real-time processing and low energy consumption, such as autonomous robots, drones or smart sensors.



Source: Basic Guide to Spiking Neural Networks for Deep Learning | Intel® Tiber™ AI StudioBasic Guide to Spiking Neural Networks for Deep Learning | Intel® Tiber™ AI Studio

SNNs are closely linked to neuromorphic chips such as Intel's Loihi or BrainChip's Akida, and are one of the AI technologies we have today that most closely resembles how the biological brain actually works.

This works more like a brain than a computer. Each artificial neuron accumulates signals over time and only sends a signal onward when a threshold is reached. This does not happen continuously, only when it is actually needed. That dramatically reduces energy consumption.

Learning happens through plasticity, particularly through mechanisms such as Spike-Timing-Dependent Plasticity (STDP). Here, the connection between two neurons is adjusted based on how close in time they send signals. The network therefore learns through patterns in timing, not just quantity.

Information in SNNs is not encoded in the strength of the signals, but in the timing and frequency of the spikes. This opens up an entirely different form of efficient and time-sensitive computation.

SNNs have several properties that make them attractive in the AI infrastructure of the future:

1. *Energy efficiency*

   They are extremely power-efficient because neurons are only active when something happens. On neuromorphic chips, this can mean up to 1,000 times lower energy consumption than traditional networks.

2. *Real-time processing*

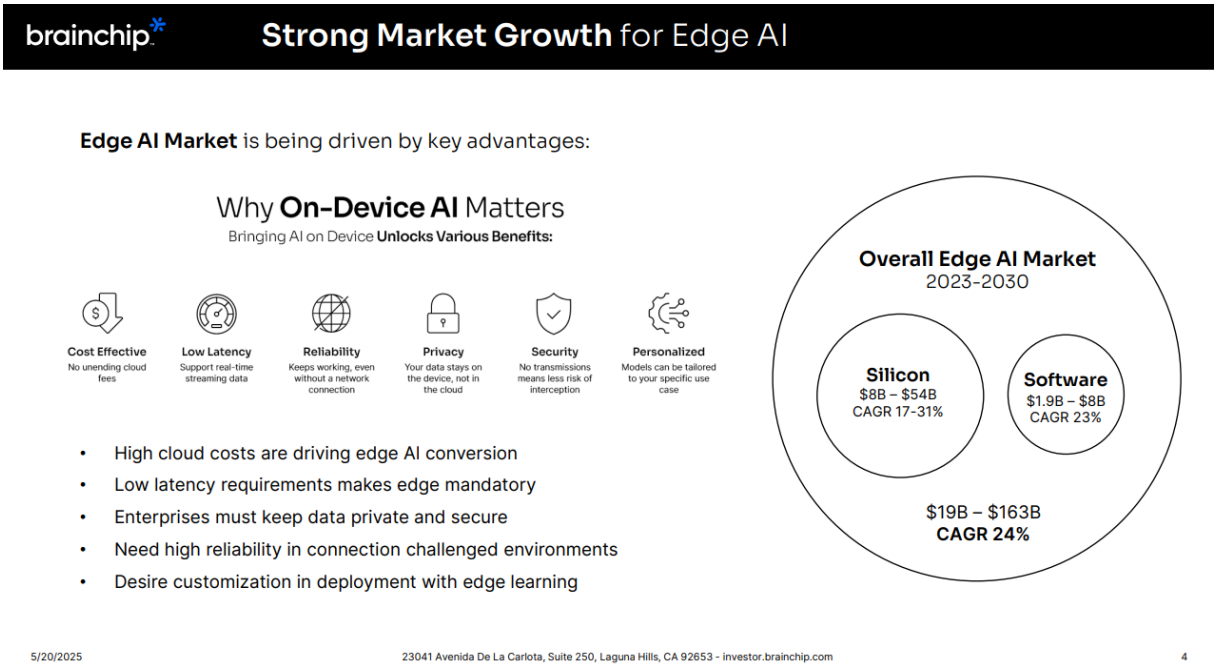
   Because they operate asynchronously and react immediately to new signals, they are well suited for applications that require fast responses, such as robotics and sensor processing.

3. *Biological plausibility*

   Because they mimic the brain's signal processing, SNNs are not only technically efficient; they are also an important tool in research on cognition and artificial general intelligence (AGI).

4. *Local learning*

   SNNs can learn directly on the device, without sending data to the cloud. This provides lower latency and better privacy.

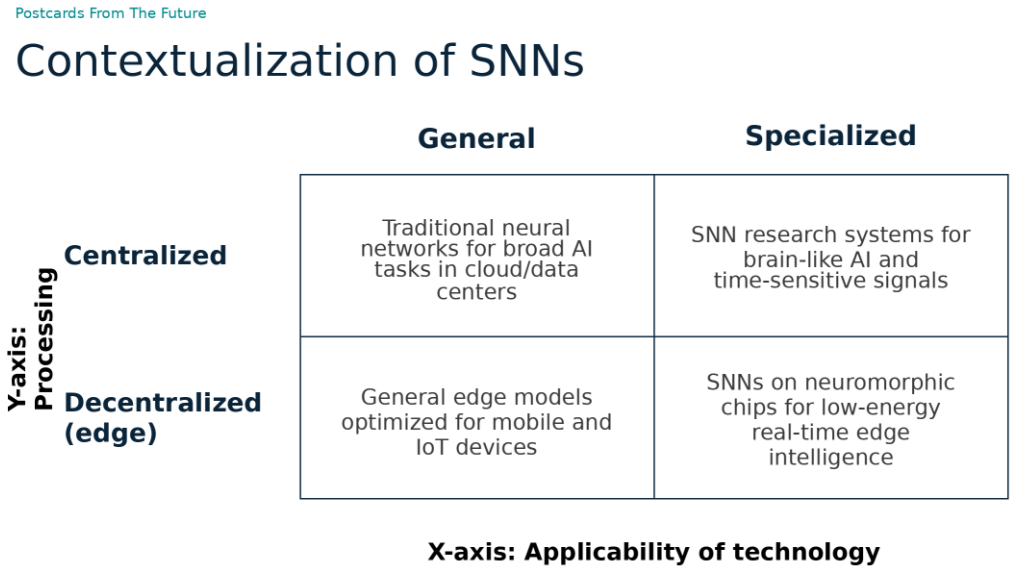


Source: BrainChip IR

However, the ecosystem is still in an early phase, with few frameworks and limited ease of use. The application area is narrow and best suited for niche tasks with time-series data. In addition, they require specialized hardware, which makes integration demanding for many.

SNNs can be understood both as part of software infrastructure, through simulation tools, and as part of compute infrastructure, when they run on neuromorphic chips. They are especially relevant in:

- Edge AI: when low energy consumption and fast response are needed, typically in mobile devices or smart sensors.
- Real-time applications: where the data stream is continuous and time-sensitive, as in autonomous vehicles or speech recognition.
- Neuroscience research: where the goal is to mimic the brain in order to better understand how intelligence works.



SNNs bring AI closer to the brain, both in how information is processed and how learning happens. They are a textbook example of biologically inspired innovation, where neuroscience meets technological development. Within AI infrastructure, they form a specialized layer of the system, optimized for time-dependent and sparse data with low power consumption. But they are still at the early-adopter stage, and both technological maturity and competence development are needed before they gain broad traction.

Some examples of AI chips:

IBM TrueNorth is a neuromorphic chip developed by IBM Research and introduced in 2014. It was a breakthrough in the pursuit of building hardware that mimics the brain, and is based on Spiking Neural Networks (SNNs). Instead of processing information with continuous values as ordinary neural networks do, TrueNorth uses discrete spikes, just like neurons in the brain.

The chip was developed as part of DARPA's SyNAPSE program and was designed with one goal in mind: extreme energy efficiency. With power consumption of only around 70 milliwatts for one million neurons, it was a response to the challenge of energy-intensive AI systems long before sustainability became a burning topic in technology.

It is built as a grid of neural cores, 4,096 in total, where each core contains 256 artificial neurons and 256 x 256 synapses. The architecture is inspired by the brain, with integrated memory and processing. This reduces the need to send data back and forth.

The chip is asynchronous and event-driven. This means processing only happens when a signal actually arrives, reducing both energy consumption and unnecessary calculations.

Learning mainly happens outside the chip, on traditional machines. In other words, TrueNorth is primarily an inference machine, optimized to use a pretrained model in an extremely power-efficient way.

What is it used for?

TrueNorth has primarily been used in research, but has also been tested in practical applications such as real-time image and audio analysis, robotics and sensor fusion. It has proven particularly effective in:

- Edge AI, where power consumption and space are limited
- Sensor processing, such as combining data from cameras, microphones and other sensors
- Neuroscience, to simulate and understand the brain's function

Compared with Loihi and Akida

All three, TrueNorth, Intel Loihi and BrainChip Akida, use SNNs and focus on low-energy processing. But they have different strengths:

	TrueNorth	**Loihi**	**Akida**
Focus	Research	Balance between research/application	Commercialization
On-chip learning	Limited	Yes (STDP and more)	Yes
Launch year	2014	2018 (Loihi 1), 2021 (Loihi 2)	2021 (Akida 1), 2023 (Akida 2)
Ecosystem	NS1e, NorthPole	Lava	Integration with Arm/RISC-V

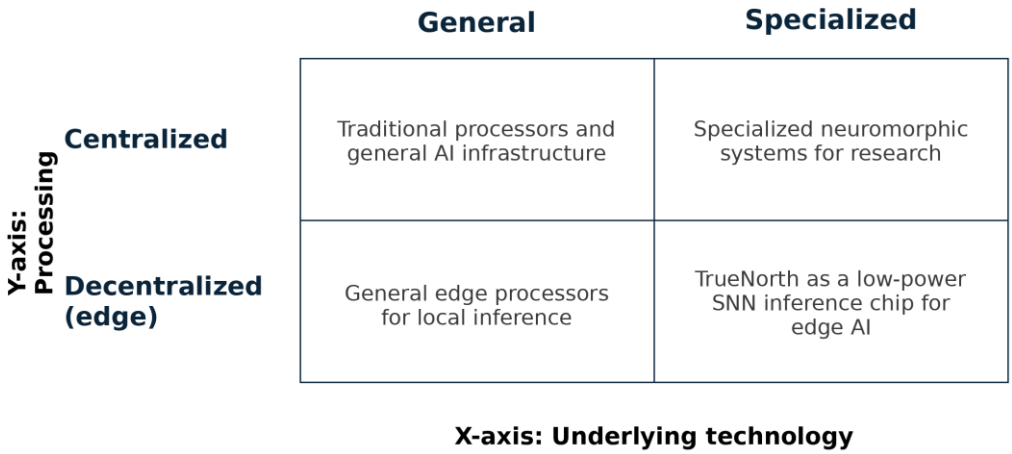
TrueNorth came first, but Loihi and Akida have since taken the technology further and developed more advanced functions and better integration with modern AI tools.

The chip was introduced at a time when AI was dominated by GPUs and cloud-based computing. It broke with the established paradigm and showed that there were other, more energy-efficient ways to build intelligence.

From 2014 onward, TrueNorth represented a shift toward specialized hardware and sustainable AI. While it has mainly been a research platform, it laid the foundation for today's and tomorrow's edge AI.

Postcards From The Future

TrueNorth in AI infrastructure



TrueNorth is a clear example of systems innovation. It challenged the prevailing architecture with a new principle: integrated memory and computation. At the same time, it shows how early-stage technology can pave the way for new paradigms, even if it does not always become the most widely used solution.

From a sustainability perspective, TrueNorth's energy efficiency is a warning signal to an industry with ever higher power consumption. Today, it has been surpassed by newer solutions such as Loihi 2 and Akida 2.

Intel Loihi is a neuromorphic chip developed by Intel Labs to mimic the brain's way of processing information. Launched in 2017 and further developed into Loihi 2 in 2021, Loihi represents one of the most ambitious attempts to create AI that works like the brain. Fast, adaptive and extremely energy-efficient.

Source: https://www.anandtech.com/show/16960/intel-loihi-2-intel-4nm-4

Unlike traditional AI chips, Loihi also uses Spiking Neural Networks (SNNs) and, like TrueNorth, operates asynchronously. With Loihi 2, Intel took the technology a step further: the chip became faster, more flexible and even more energy-efficient. In 2024, Intel launched Hala Point, the world's largest neuromorphic system, with more than one billion neurons, built on Loihi 2, to show how the technology can scale for large compute tasks.

Loihi is developed with the AI systems of the future in mind, and is already used in research and development of edge AI. Relevant applications include:

Comparison:

	Intel Loihi	**BrainChip Akida**
Focus	Research and scalability	Commercial edge AI
On-chip learning	Yes (STDP and adaptive learning)	Yes (also STDP-based)
Scalability	Loihi 2 and Hala Point	Compact design for IoT
Ecosystem	Lava (open source)	Arm/RISC-V integration

Although both use SNNs, Loihi is more research-oriented and scalable, while Akida focuses on practical and commercial applications.

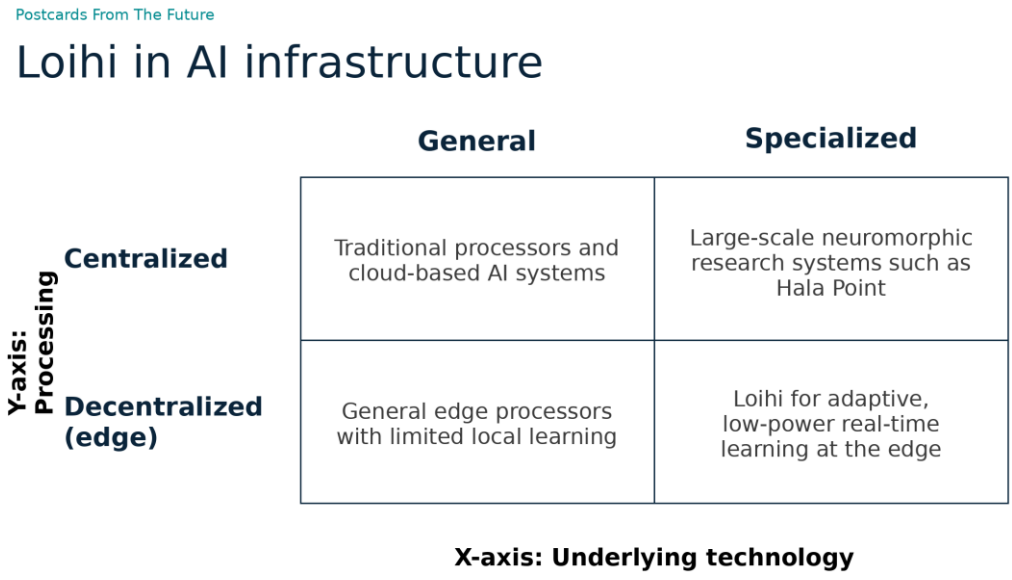
GPUs and cloud-based solutions are still dominant, but Loihi points toward a future where AI moves closer to the data source, out to sensors, drones and devices where power and latency mean everything.

Advantages:

- Highly energy-efficient
- Supports real-time learning and adaptive algorithms
- Suitable for edge AI where cloud connectivity is limited
- Can be scaled to large-scale systems (Hala Point)

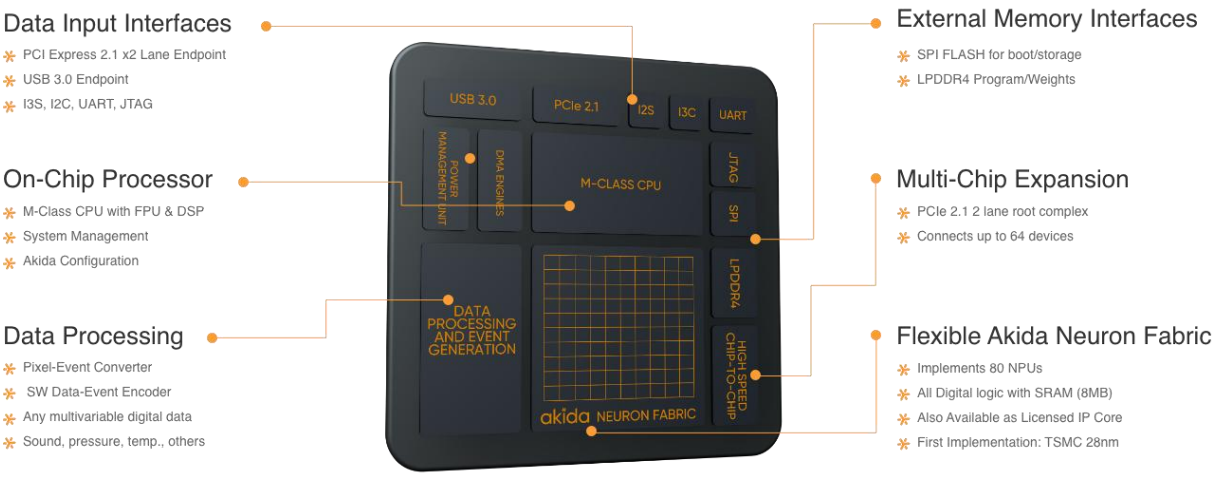
Limitations:

- Currently mostly used in research
- The ecosystem is still under development compared with TensorFlow
- Not as well suited for heavy language and image tasks
- Limited commercial adoption; still in an early phase



Through its biologically inspired approach, it addresses two of the AI industry's biggest challenges: energy consumption and latency. With support for local learning and the open-

source platform Lava, Loihi invites researchers and developers to experiment and build new models for brain-like AI. At the same time, the technology is still early in the innovation adoption curve, but its potential to shape the AI infrastructure of the future is significant.



BrainChip "Akida". Source: BrainChip.com

BrainChip is an Australian technology company that builds AI processors inspired by the brain. Its flagship product, Akida, is a neuromorphic chip designed to make artificial intelligence extremely energy-efficient and run directly on the device, without cloud connectivity. This is called edge AI, and it is where BrainChip tries to differentiate itself.

The company was founded in 2004, but it was not until the launch of Akida in 2018 and the second generation in 2023 that it truly became a relevant actor in AI infrastructure. Its technology is based on Spiking Neural Networks (SNNs), where information is processed as spikes, just as biological neurons do. This provides both low power consumption and the possibility of fast, local learning.

What makes Akida unique?

Unlike traditional AI chips that require continuous cloud access, Akida allows the device to learn and adapt locally, for example using learning principles such as STDP (Spike-Timing-Dependent Plasticity). This provides major advantages in applications where privacy, real time and power saving are important, such as health sensors, autonomous vehicles and industrial IoT systems.

BrainChip collaborates with major actors such as Arm and Andes Technology, and Akida can be integrated with both RISC-V and Cortex-M processors. This makes the technology easy to use for developers already working with edge devices.

BrainChip represents a decentralized alternative to today's dominant AI models, which require powerful GPUs and data centers. Instead of sending all data to the cloud, Akida brings intelligence all the way out to the edge, close to the sensor or the user. This can save energy, reduce latency and operate without constant network connectivity.

Important milestones

- 2004: BrainChip was founded in Perth, Australia.
- 2018: the first generation of Akida is launched.
- 2023: Akida 2 is launched, with support for Vision Transformers and spatiotemporal AI.
- 2025: expanded integrations with RISC-V and participation at major industry conferences.

Advantages:

- Extremely energy-efficient, well suited for battery-powered devices
- Real-time processing and adaptation without cloud connectivity
- Integrates easily into existing ecosystems through collaboration with Arm and Andes
- Promotes privacy because data does not have to be sent out of the device

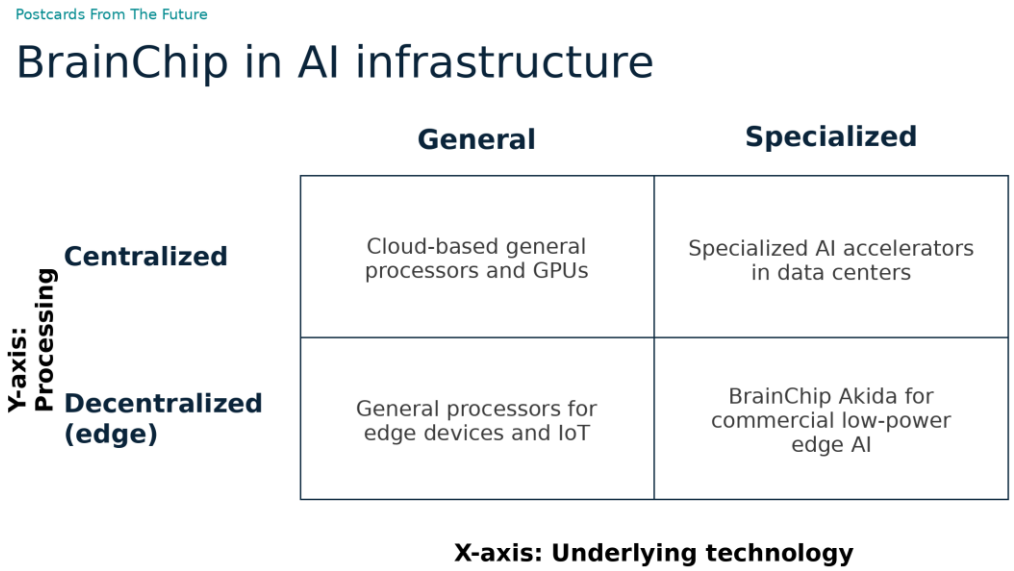
Challenges:

- Neuromorphic technology is still in an early phase
- Lacks the same software support as TensorFlow and PyTorch
- Limited to specific use cases; not suitable for large language or image-based models
- Competes against giants such as Intel and NVIDIA, which have larger resources and networks

Comparison: BrainChip Akida vs. Intel Loihi

	BrainChip Akida	Intel Loihi
Main focus	Commercial edge devices	Research and scalability
On-chip learning	Yes, via STDP	Yes, via STDP and other mechanisms
Typical applications	IoT, cars, sensors	Robots, neuroscience, labs
Ecosystem	Arm, RISC-V integrations	Lava (open source from Intel)
Maturity	More commercial	More research-oriented

BrainChip builds on disruptive innovation, where smaller and specialized actors introduce technology that initially seems niche, but can later take over markets when needs change. This is especially relevant in smart cities, autonomous transport systems and IoT, areas where energy efficiency and the edge are crucial.



BrainChip is a good example of how systems theory and innovation theory meet in practice. Akida is not designed to replace all AI semiconductors, but to fill a specialized need and do it very efficiently. The technology challenges established paradigms for how and where AI should be trained and run, and shows how edge AI and neuromorphic systems can become an important part of a sustainable and privacy-friendly future.

Tesla and NVIDIA

Tesla is more than just a car manufacturer. It is a technology company that develops autonomous vehicles through FSD, humanoids through Optimus, and energy systems powered by artificial intelligence. At the core of this effort is the company's ambition to build a fully vertically integrated AI infrastructure, for both training and inference.

NVIDIA is the global leader in AI hardware and software, known for its GPUs, SoCs (system on chip) and tools such as CUDA and TensorRT. NVIDIA is the backbone of modern AI infrastructure, both in the cloud and at the edge.

Tesla and NVIDIA collaborate closely, but also compete directly. This dynamic plays out especially in two central parts of the AI lifecycle: training (model development) and action (inference).

Training: interaction and independence

Collaboration

- **Hardware: Tesla uses NVIDIA GPUs (H100, A100) in data centers to train FSD and Optimus.**
- **Software: CUDA and AI Enterprise give Tesla powerful tools for optimized training.**

Competition

- **Hardware: Tesla has developed Dojo, a supercomputer with proprietary D1 chips.**
- **Software: Tesla builds its own MLOps stack and training tools, optimized for Dojo.**

Seen through the lens of innovation theory, NVIDIA disrupted CPU dominance, while Tesla is attempting a counter-disruption with Dojo.

Action (Inference): edge in motion

Collaboration

- **Hardware: Tesla uses NVIDIA DRIVE Orin in Model S/X for real-time analysis and navigation.**
- **Software: NVIDIA's TensorRT and DRIVE OS make model execution in the cars more efficient.**

Competition

- **Hardware: Tesla develops its own SoCs (HW3/HW4), which replace DRIVE Orin.**
- **Software: Tesla's inference systems are adapted to its own hardware and data streams.**

Systems theory: NVIDIA functions as a support system for Tesla, but Tesla is simultaneously building an independent ecosystem.

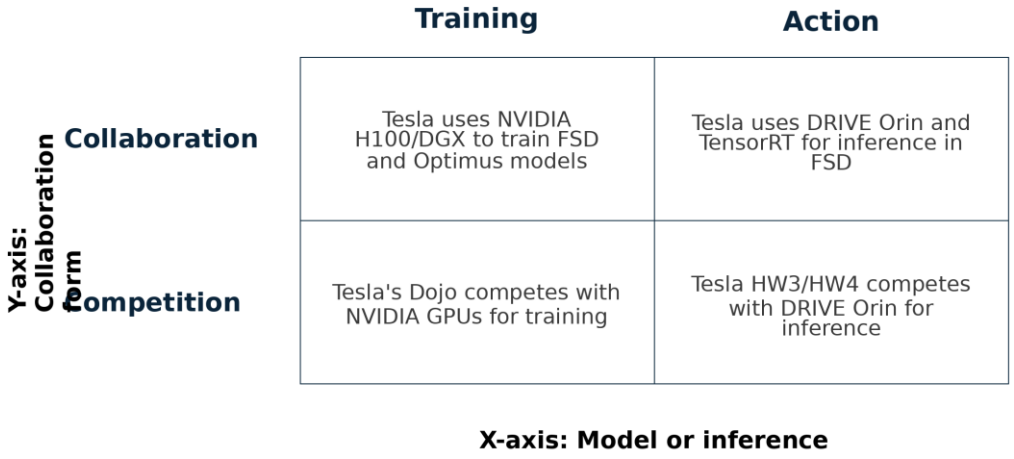
Tesla collects and uses data directly from millions of cars to improve its models. Federated learning makes it possible to update the models without sending raw data to the cloud.

- Collaboration: NVIDIA's platforms support federated learning and local processing.
- Competition: Tesla optimizes HW3/HW4 for local learning and direct communication with Dojo.

Tesla gains an advantage by minimizing cloud dependence and protecting user data.

Postcards From The Future

Tesla and NVIDIA in training and action



Tesla competes against	NVIDIA competes against
NVIDIA (GPUs, DRIVE)	AMD, Google TPU, Intel Gaudi, Cerebras
Google, Amazon (Trainium)	Tesla HW3/HW4, BrainChip, Qualcomm
Qualcomm, Mobileye	Intel Loihi, Google Edge TPU

Strengths

Tesla has full control of the AI stack, an enormous fleet with real-time data and tailored hardware and software

NVIDIA with strong developer lock-in through CUDA, market leader in AI training, and an ecosystem with a broad customer base

Disruption examples

Tesla	NVIDIA
Dojo challenges GPU dominance	GPUs disrupted CPU training (2006)
HW3/HW4 vs. DRIVE Orin	DRIVE disrupted cloud-based inference
Federated Learning replaces centralized data flow	Redefined data centers as AI factories

Postcards From The Future

Tesla and NVIDIA in training and action

Y-axis: Collaboration vs Competition	Training	Action
Collaboration	Tesla uses NVIDIA H100/DGX to train FSD and Optimus models	Tesla uses DRIVE Orin and TensorRT for inference in FSD
Competition	Tesla's Dojo competes with NVIDIA GPUs for training	Tesla HW3/HW4 competes with DRIVE Orin for inference

X-axis: Model or inference

- Systems theory: NVIDIA delivers foundational infrastructure; Tesla builds specific subsystems.
- Innovation theory: interaction between cooperative and disruptive innovation.
- Privacy theory: Tesla creates value by keeping and processing data locally.
- Sustainability theory: both focus on lower energy consumption, but Tesla goes further with specialized chips.

NVIDIA’s role in training and action in artificial intelligence

Jensen Huang has taken NVIDIA from being a producer of graphics cards to becoming a cornerstone of the future economy. Not only does the company deliver the world’s most sought-after AI chips, it has also built an entire ecosystem, a kind of operating system for intelligence, stretching from data centers to self-driving cars. NVIDIA is no longer just a company that sells chips. It is a provider of infrastructure for artificial intelligence.

1. Training:

AI models must be trained, and training takes place in data centers or “AI factories” that convert raw data and electricity into intelligence. Here, NVIDIA dominates completely.

Hardware:
Its GPUs (H100, A100 and now Blackwell B200) are designed for massive parallel processing, making them ideal for training large language models and visual models. DGX systems, which combine multiple GPUs with storage and networking, are used by companies such as OpenAI, Meta and Tesla.

Software:
CUDA, NVIDIA’s programming framework, is an industry standard. Together with tools such as Triton Inference Server, RAPIDS and Omniverse, NVIDIA gives developers and researchers tools for data processing, model training and simulation.

Customers:
OpenAI trains GPT-4 with NVIDIA. Tesla uses them for FSD and Optimus. Google, Microsoft and Amazon offer them in the cloud. NVIDIA is the infrastructure itself.

Competitors:
AMD challenges with MI300X, but lacks NVIDIA’s ecosystem. Google has TPUs, but they are less flexible. Startups such as Cerebras and Groq are innovative, but small.

Disruption:
NVIDIA replaced CPUs as the training engine with GPUs and defined a new industry standard

with CUDA. Now the company is moving further with AI factories and synthetic data from Omniverse.

2. Action:

After training, the models must go out and do something. This is called inference. It can mean interpreting a patient record, navigating a robot or driving a car. Here, NVIDIA offers both cloud-based and edge solutions.

Hardware:
Jetson Orin and DRIVE Orin are specialized system-on-chips (SoCs) for Edge AI, used in robotics and self-driving vehicles. In the cloud, H100 and A100 are also used for inference in applications such as chatbots and search.

Software:
TensorRT optimizes models for low latency and energy efficiency. NVIDIA AI Enterprise and Triton are used for inference in the cloud. Isaac Sim and DRIVE OS are used for robotics and autonomous driving.

Customers:
Waymo, Mercedes and Tesla use DRIVE for self-driving cars. AWS offers NVIDIA GPUs in its cloud services. Jetson is used in everything from robots to smart cameras.

Competitors:
Qualcomm, Intel and BrainChip challenge with low-energy edge chips. Google has Edge TPU. But NVIDIA still has the strongest integration between hardware and software.

Postcards From The Future

NVIDIA's role in training and action

Y-axis: Processing	Training	Action
Centralized	NVIDIA dominates AI training in cloud and data centers with GPUs	NVIDIA enables cloud inference for chatbots, search and analytics
Decentralized (edge)	Federated learning via DRIVE lets vehicles learn locally and send insights back	Jetson and DRIVE enable real-time inference in robots and cars

X-axis: Model training or inference

Explanation:

- *Training + centralized: this is where NVIDIA dominates. GPT-4 and LLaMA are trained on NVIDIA in the cloud.*
- *Training + decentralized: federated learning through DRIVE lets cars learn locally and send insights back.*
- *Action + centralized: apps such as ChatGPT or Amazon Alexa use NVIDIA in the cloud.*
- *Action + decentralized: NVIDIA enables real-time inference in robotics and the automotive industry with Jetson and DRIVE.*

Systems theory shows NVIDIA as a central subsystem in the AI ecosystem. It connects training and action, cloud and edge, through an integrated system of hardware and software. Innovation theory explains how NVIDIA has disrupted old technology models and created new standards. With sustainable data centers and support for federated learning in cars, NVIDIA also plays a key role in the development of responsible AI.

SNNs and the rise of humanoids

SNNs are ideally suited for humanoids such as Tesla Optimus, Boston Dynamics' Atlas and Figure AI's models.

While traditional AI models require enormous amounts of energy and computing power, SNNs are designed for energy-efficient real-time processing directly on the device. This is what we call Edge AI. Humanoids are often battery-powered and must react quickly to unpredictable situations. SNNs enable this, especially when they run on neuromorphic chips such as Intel Loihi or BrainChip Akida.

How are SNNs used in humanoids?

Humanoids require the ability to process visual, auditory and tactile information simultaneously, and to act adaptively and quickly. Here, SNNs offer several advantages:

Sensor fusion:
SNNs handle input from cameras, LIDAR and tactile sensors in real time, enabling the robot to perceive and react to its environment with high precision while keeping energy consumption low.

Motor control:
Complex movements such as balance, walking and gripping are controlled with high responsiveness and low latency, inspired by how the brain controls muscles.

Social interaction:
SNNs process facial expressions, speech and body language to enable more natural interaction with humans, and can adapt to different users over time.

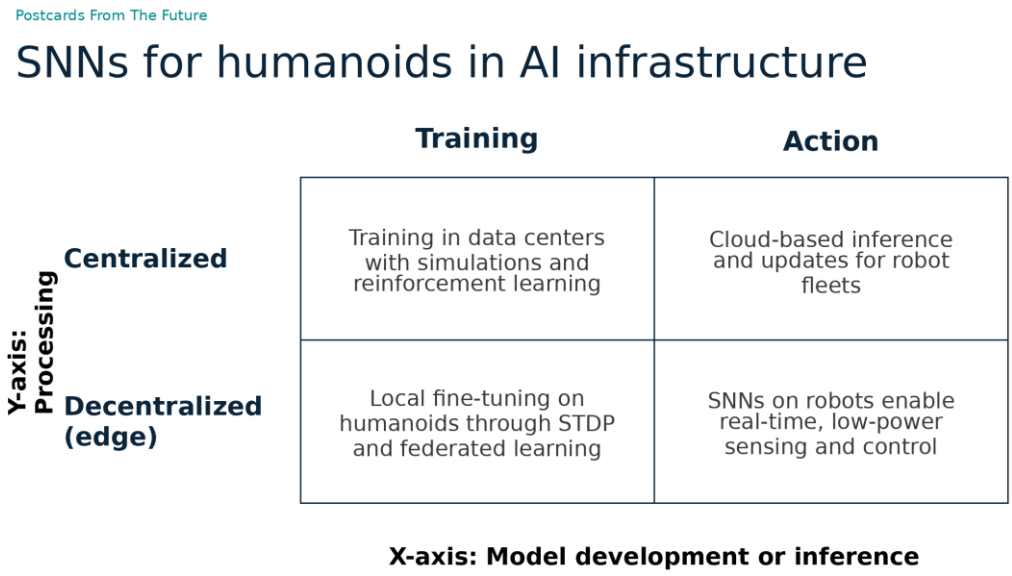
Local learning:
Through mechanisms such as Spike-Timing-Dependent Plasticity (STDP), the robot can learn and adjust behavior locally, without being connected to a central server.

Training of SNNs often begins centrally in data centers. Methods such as conversion from traditional DNNs, direct spike-based training or reinforcement learning in simulations such as NVIDIA Isaac Sim are used. Afterwards, the models are fine-tuned locally on the robot through STDP or federated learning, where each robot learns locally and shares updates without sharing personal data.

Edge adaptation requires low latency, robustness and high energy efficiency. This is where neuromorphic chips stand out: they typically use only milliwatts and enable responses within milliseconds. This is crucial for avoiding falls, navigating uneven terrain or interpreting a hand movement in real time.

When humanoids learn locally, they must also be able to store this knowledge securely and efficiently. This often happens on embedded chips or in encrypted storage, and can include everything from individual gait patterns to user preferences in conversations.

An important principle is that this learning can also be fed back into centralized models through federated learning or by anonymizing and aggregating data. In this way, collective intelligence develops based on the experiences of thousands of humanoids, without sensitive information leaving the devices.



SNNs represent a biologically inspired approach to intelligence, tailored for the robots of the future. Within systems theory, they function as specialized components that connect sensing, learning and action in an energy-efficient and autonomous way. From an innovation theory perspective, we are in an early adoption phase, but with the potential to redefine how Edge AI is performed. At the same time, sustainability theory highlights the benefits of low energy consumption, while privacy theory emphasizes the value of local learning without data sharing.

SNNs and neuromorphic chips allow humanoids not only to exist in our world, but to learn in it, from it and adapt to us.

From cloud to edge:

Whether we are talking about humanoids, AI agents or self-driving cars, it ultimately comes down to one thing: learning. To see, understand, act and improve over time. But even though these systems share common principles in machine learning and deep learning, their training models are tailored to different realities. Here, we look more closely at how models are trained in the cloud, how they are moved out to the edge, and how learning from there flows back and makes the next generation even smarter.

Humanoids:

Humanoids are robots with a physical presence, designed to mimic human movement and interaction. To master such skills, neural networks (such as SNNs and transformers) are combined with reinforcement learning (RL) and imitation learning. Movement, balance and social interpretation are all learned in simulations such as MuJoCo and Isaac Sim, where mistakes cost no more than a few milliseconds.

Several companies train humanoids on multimodal input (image, sound, touch) to give them more natural and adaptive behavior. After training in the cloud, the models are transferred to the edge, where they must function with low energy use, in real time and often without connectivity.

AI agents:

AI agents are digital, often language-based systems that operate in the browser, on mobile or in the cloud. They are based on transformer architectures and are trained as Large Language Models (LLMs), with billions of parameters and petabytes of text. After pretraining, they are fine-tuned with RLHF, Reinforcement Learning from Human Feedback, so they can respond more nuancedly and efficiently.

The largest training runs take place in the cloud on thousands of H100 GPUs. But lighter versions of agents are increasingly being moved to the edge, for example on smartphones, where they can adapt to the user without sending data back to the cloud.

Self-driving cars:

Self-driving cars operate in a physical, unpredictable world. To understand surroundings and make decisions in real time, they use neural networks such as SNNs and transformers, together with sensor fusion and RL. Tesla, Waymo and Cruise train their models in the cloud, often using simulations and data from real driving.

But for inference, meaning the actual decision-making while driving, everything happens at the edge. The car must interpret, react and learn without waiting for a cloud response. The models are therefore optimized for low latency, robustness and safety, and are often connected to systems for continuous local learning.

When intelligence moves into products

To move intelligence from the cloud to edge devices, models must be made lighter and more efficient. This means they must be simplified and trimmed down. Large language models need simpler variants, and both SNNs and transformers must be adapted to run quickly and power-efficiently on specialized chips such as NVIDIA Jetson, Qualcomm Snapdragon or neuromorphic chips such as BrainChip Akida and Intel Loihi.

Edge devices have limited power, compute capacity and memory. But they have one advantage: they are close to reality. That makes them the most valuable learning arenas we

have. Through federated learning, updates, not raw data, are sent back to the cloud. There, they are combined with the experiences of other devices to update the global model.

Tesla does this with Autopilot. Google does it with keyboard models. Apple does it with Siri. It is a new feedback loop: the cloud trains the edge, the edge learns from reality and improves the cloud.

Edge devices do not only learn in general. They learn you. Your voice, your movements and your environment. This is called locally adapted action. It often happens through on-device learning, STDP or light fine-tuning, also called LoRA fine-tuning.

In humanoids, this means the robot learns your way of pointing. In AI agents, it means the agent understands how you express yourself. In cars, it means they adapt to the roads you use every day.

Everything happens locally. Everything is stored locally.

A new feedback loop

Training and action are no longer separated. They form a loop: models are trained in the cloud, distributed to the edge, learn in real time and improve the central model through feedback. The result is a network of intelligent nodes that learn collectively.

The next time you see a robot walk, a car turn by itself or an AI complete your sentence, remember that you are part of the learning. It is not only the machines that are evolving. It is the infrastructure around us.

Where are the largest AI opportunities in 2025?

For those of us who follow emerging technology, the question is no longer whether artificial intelligence will change industries, but where and how. Based on reports and analyses from McKinsey and Gartner, among others, we point to seven main categories where AI creates economic growth, margin improvements and scalability. This chapter summarizes these

opportunities, connects them to theoretical frameworks and identifies listed companies that give investors concrete exposure to the development.

This is meant only as information and should not be considered investment advice. Always do your own analysis.

Some opportunities in 2025

1. **AI-driven automation of business processes**
   Artificial intelligence streamlines repetitive tasks such as customer service, accounting and logistics. Typical solutions include RPA and AI chatbots. This provides high scalability and low variable costs, with very high gross margins.
2. **Generative AI in creative services**
   AI that generates text, images, video and music enables value creation in marketing, entertainment and game development. Tools such as ChatGPT and DALL-E reduce production costs and open up new business models.
3. **Cybersecurity**
   As cyber threats become more complex, AI plays a crucial role in real-time threat detection and response. Many actors here have chosen SaaS models with high margins and subscriptions.
4. **Healthcare and biotechnology**
   AI makes it possible to analyze medical images, genomic data and symptom patterns with high precision. This lowers costs, improves diagnostics and accelerates drug development.
5. **Semiconductors and infrastructure**
   Under the hood of large language models and AI agents are specialized AI chips and data centers. This is the most capital-intensive, but also the most indispensable, part of the AI ecosystem.
6. **Sustainability and energy optimization**
   From smart grids to greener data centers, AI enables energy efficiency and emissions reductions. This happens alongside regulatory requirements and the growing need for stable, sustainable energy use.

7. **Education and learning technology (EdTech)**
   AI-driven learning adapted to each individual user is growing rapidly. Products such as Duolingo show how scalable, digital solutions can create high user engagement and profitability in learning.

These points are not random. From a classical economic perspective, they build on economies of scale and intangible assets, which enable high returns on capital. In financial analysis methodology, gross margin, revenue growth and the Rule of 40 are especially important in identifying actors that combine growth and profitability. And from innovation theory, we see that many of these categories are in the exponential phase of the S-curve, with rapid adoption and significant value creation.

We look for companies with high margins, good growth and possible mispricing. This leads us to a range of listed companies that provide direct or indirect exposure to AI trends:

- **UiPath and ServiceNow (automation)**
- **Unity Software and Adobe (generative AI)**
- **Zscaler and SentinelOne (cybersecurity)**
- **Tempus and Intuitive Surgical (health technology)**
- **NVIDIA and AMD (infrastructure and AI chips)**
- **NextEra Energy and Schneider Electric (sustainability/energy)**
- **Duolingo and Coursera (AI in education)**

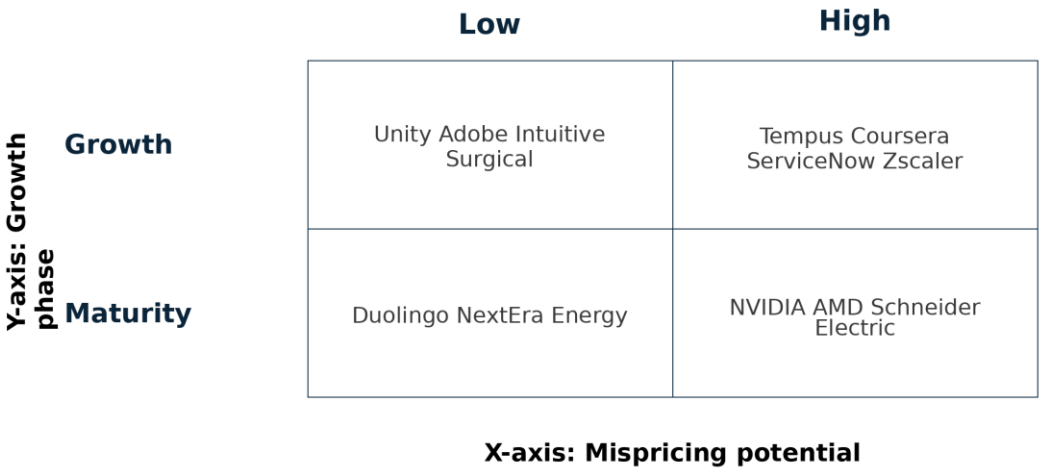
Several of these companies have high beta and are affected by negative market sentiment, but combine that with strong growth and profitability. This makes them particularly interesting for a contrarian approach, where the market underestimates future earnings.

To understand where the opportunities are greatest, we have placed the companies in a 2x2 matrix based on two dimensions:

1. S-curve: is the company in an exponential growth phase or in a mature phase?
2. Mispricing potential: is the market negatively positioned despite strong fundamentals?

Postcards From The Future

Companies within AI



Companies in the upper-right quadrant combine high growth and high margins with low valuation and negative sentiment. This is typically where the market overreacts to short-term problems. This applies, for example, to Zscaler and Tempus. In the lower-left quadrant, we find solid companies with good margins and low beta, but also lower expected upside, such as Intuitive Surgical or Adobe.

The seven categories all have in common that they combine high scalability with the potential for profitability in an increasingly automated economy. This provides a valuable framework for identifying companies with high margins, growth and the potential for mispricing. In a deflationary, robotized economy where productivity and software gain ground, these actors may become among the most important growth engines of the future.